Fig. 13. Exact numerical results for the repository size required to detect association is shown as a function of the allele frequency p for (A) dominant inheritance, (B) additive inheritance, and (C) recessive inheritance for tests using pooled DNA. The variance ratio σ_A^2/σ_R^2 is 0.02, the type I error is 5×10^{-8}, the type II error is 0.2, the pooling fraction 0.27 is used for all designs except Mahalanobis, for which 0.188 is used. The Mahalanobis design loses power for rare alleles faster than the other designs.

Fig. 14. Exact numerical results for the repository size required to detect association is shown as a function of the heterozygote phenotypic displacement d, describing the inheritance mode, for allele frequencies of (A) $p = 0.5$, (B) $p = 0.25$, and (C) $p = 0.1$ for tests using pooled DNA. All other parameters are as in Fig. 13.

Fig. 15 The repository size required to detect association for a QTL for a complex trait is shown for pooled DNA designs relative to individual genotyping designs having equivalent type I and type II error rates. The ratio $N_{aff/unaff}/N_{indiv}$ for affected/unaffected pools (dashed line) is shown as a function the disease prevalence r, while the ratio N_{tail}/N_{indiv} (solid line) is shown as a function of the fraction ρ of the total population selected for each pool. The optimum value of N_{tail}/N_{indiv} is 1.24 and occurs at $\rho = 27.03\%$ selected for each pool.

Fig. 16 The effect of varying the inheritance mode is shown for tail pools. The type I error is 5×10^{-8}, the type II error rate is 0.2, and the displacement a is 0.25 in units of the phenotypic standard deviation. The displacement d of heterozygotes varies from $-a$, pure recessive inheritance, to $+a$, pure dominant inheritance. Three allele frequencies are shown, $p = 0.5$, 0.1, and 0.01. Solid lines correspond to exact numerical calculations. (Top) The repository size N is shown. Filled circles corresponding to analytical approximations, Eq. 1, are virtually indistinguishable from exact calculations. (Bottom) The optimal pooling fraction ρ from numerical calculations falls in a narrow range from 24.5% to 27.5%, close to the analytical approximation of 27.03%.

Fig. 17 (Top) Exact numerical results for the repository size N required to achieve a type I error rate of 5×10^{-8} and type II error rate of 0.2 are shown for affected/unaffected pools (dashed line) and tail pools (solid line) as a function of the additive variance, also presented as the genotype relative risk for a heterozygote, for an allele with frequency 0.1 and purely additive inheritance. Analytical approximations (solid circles), Eqs. 1 and 2, are indistinguishable from the exact results when the genotype relative risk is smaller than 2. The disease prevalence r is 10% for the affected/unaffected pools, and 27% of the population is selected for each of the tail pools. (Bottom) The frequency difference at the significance threshold is shown for the same parameters. This threshold determines the measurement accuracy required for association tests based on pooled DNA.

Detailed Description of the Invention

1. Definitions

Glossary of mathematical symbols

X	quantitative phenotypic value of an individual
X_i	quantitative phenotypic value of sib i, with $i = 1$ or 2 for sib-pairs
$X_\pm$	$(X_1 \pm X_2)/2$
r	phenotypic correlation between sibs
A_i	allele inherited at a particular locus. For a bi-allelic marker, $i = 1$ or 2
G	genotype at the locus, either A_1A_1, A_1A_2, or A_2A_2 for a bi-allelic marker
G_i	genotype for sib i, with $i = 1$ or 2 for sib-pairs
$P(G)$	genotype probability
$P(G_1,G_2)$	joint sib-pair genotype probability
$f(X_1,X_2)$	joint sib-pair phenotype probability distribution
$f[X_1,X_2\|G_1,G_2]$	joint sib-pair phenotype probability distribution conditioned on genotypes
p	frequency of allele A_1 in a population
q	frequency of the remaining alleles, with $q = 1 - p$
p_i	frequency of allele A_1 in sib i, either 1, 0.5, or 0 for an autosomal marker

$p_{\pm}$ $(p_1 \pm p_2)/2$

a half the difference in the shift in the mean phenotypic value of individuals with genotype A_1A_1 compared to A_2A_2

d difference in the mean phenotypic value between individuals with genotype A_1A_2 compared to the mid-point of the means for A_1A_1 and A_2A_2

μ mean phenotypic shift due to the locus, equal to $a(p-q) + 2pqd$

σ_A^2 additive variance of phenotype X due to the genotype G

σ_D^2 dominance variance due to the genotype G

σ_R^2 residual phenotypic variance, with $\sigma_A^2 + \sigma_D^2 + \sigma_R^2 = 1$

N the total number of individuals whose DNA is available for pooling

n number of individuals selected for a single pool

ρ pooling fraction defined as n/N

p_U, p_L frequency of allele A_1 in the upper (U) or lower (L) pool

T test statistic, which is expected to be close to zero when the genotype G does not affect the phenotypic value and is expected to be non-zero when individuals with genotypes A_1A_1, A_1A_2, and A_2A_2 have different mean phenotypic values. As formulated here, T has a normal distribution with unit variance. Under the null hypothesis that $\sigma_A = (2pq)^{1/2}[a-(p-q)d]$ is zero, the mean of T is zero. Under the alternative hypothesis that σ_A is non-zero, the mean of T is also non-zero.

σ_0^2 variance of $n^{1/2}(p_U-p_L)$ under the null hypothesis

σ_1^2 variance of $n^{1/2}(p_U-p_L)$ under the alternative hypothesis

$\Phi(z)$ cumulative standard normal probability, the area under a standard normal distribution up to normal deviate z

z_α normal deviate corresponding to an upper tail area of α, defined as $\Phi(z_\alpha) = 1-\alpha$

α type I error rate (false-positive rate). For a one-sided test, $T > z_\alpha$ corresponds to statistical significance at level α, typically termed a p-value. A typical threshold for significance is a p-value smaller than 0.05 or 0.01. If M independent tests are conducted, a conservative correction that yields a final p-value of α is to use a p-value of α/M for each of the M tests.

β type II error rate (false-negative rate). The power of a test is $1-\beta$.

$H(x)$ Heaviside step function

As used herein, when two individuals are "related to each other", they are genetically related in a direct parent-child relationship or a sibling relationship. In a sibling relationship, the two individuals of the sibling pair have the same biological father and the same biological mother. As used herein, the term "sib" is used to designate the word "sibling", and the sibling relationship is defined above. The term "sib pair" is used to designate a set of two siblings.

The members of a sib pair may be dizygotic, indicating that they originate from different fertilized ova. A sib pair includes dizygotic twins.

The focus of the present invention is to examine the statistical power of pooling designs for quantitative phenotypes. A variance components model provides the distribution of phenotypic values for an unselected population of unrelated individuals or sib pairs. The phenotype is partitioned into contributions from a specific causative allele and from residual shared and non-shared familial and genetic factors. The genotype-dependent phenotype distribution for sib pairs under Hardy-Weinberg equilibrium is used as the basis for analyzing the statistical power of various pooling strategies. The test statistic in each case is the allele frequency difference between two pools, appropriately standardized to a normal distribution. Numerically exact results are provided for a range of parameters including the fraction of population pooled, the allele frequency, and the dominant or recessive character of the allele. Furthermore, upon consideration of the relative powers of pooling designs, pooling designs are suggested for particular phenotype characteristics.

2. Model 1

2.1 Biometrical Genetic Model

A quantitative phenotype X, standardized to zero mean and unit variance, is hypothesized to be affected by the genotype G at a biallelic locus with alleles A_1 and A_2, occurring at population frequencies p_1 and $p_2 = 1-p_1$. More generally, A_2 may represent any of a number of alternate alleles, and p_2 their aggregate frequency. The population is assumed to be random mating, with genotype frequencies $P(G)$ of p_1p_1, $2p_1p_2$, and p_2p_2 for A_1A_1, A_1A_2, and A_2A_2 respectively. The frequency of allele p_1 in genotype G, denoted p_G, is 1 for A_1A_1, 0.5 for A_1A_2,

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

B1

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 February 2002 (28.02.2002)

PCT

(10) International Publication Number
WO 02/16643 A2

(51) International Patent Classification[7]: C12Q 1/68

(21) International Application Number: PCT/US01/25924

(22) International Filing Date: 20 August 2001 (20.08.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/226,465	18 August 2000 (18.08.2000)	US
60/230,580	5 September 2000 (05.09.2000)	US
09/932,480	17 August 2001 (17.08.2001)	US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier applications:

US	60/226,465 (CIP)
Filed on	18 August 2000 (18.08.2000)
US	60/230,580 (CIP)
Filed on	5 September 2000 (05.09.2000)
US	09/932,480 (CIP)
Filed on	17 August 2001 (17.08.2001)

(71) Applicant *(for all designated States except US)*: **CURAGEN CORPORATION** [US/US]; 555 Long Warf Drive, 11th floor, New Haven, CT 06511 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **BADER, Joel, S.** [US/US]; 1127 High Ridge Road, #107, Stamford, CT 06905 (US). **BANSAL, Aruna** [US/US]; 322 East Main Street, Branford, CT 06405 (US). **SHAM, Pak** [US/US]; 322 East Main Street, Branford, CT 06405 (US).

(74) Agent: **ELRIFI, Ivor, R.**; Mintz, Levin, Cohn, Ferris Glovsky and Popeo, P.C., One Financial Center, Boston, MA 02111 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: DNA POOLING METHODS FOR QUANTITATIVE TRAITS USING UNRELATED POPULATIONS OR SIB PAIRS

(57) Abstract: Identifying the genetic determinants for disease and disease prediposition remains one of the outstanding goals of the human genome project. When large patient populations are available, genetic approaches using single nucleotide polymorphism markers have the potential to identify relevant genes directly. While indivieual genotyping is the most powerful method for establishing association, determining allele frequencies in DNA pooled on the basis of phenotypic value can also reveal association at much-reduced cost. Here we analyze pooling methods to establish association between a genetic polymorphism and a quantitative phenotype. Exact results are provided for the statistical power for a number of pooling designs where the phenotype is described by a variance components model and the fraction of the population pooled is optimized to minimize the population requirements. For low to moderate sibling phenotypic correlation, unrelated population requirements. For low to moderate sibling phenotypic correlation, unrelated populations are more powerful than sib pair populations with an equal number of individuals, for sibling phenotypic correlations above 75 %, however, designs selecting the sib pairs with the greatest phenotype difference become more powerful. For sibling phenotype correlations below 75 %, pooling extreme unrelated individuals is the most powerful design for sib pair populations. The optimal pooling fractions for each design are constant over a wide range of parameters. These results for quantitative phenotypes differ from those reported for qualitative phenotypes, for which unrelated populations are more powerful than sib pairs and concordant designs are more powerful than discordant, and have immediate relevance to ongoing association studies and anticipated whole-genome scans.

WO 02/16643 A2

DNA Pooling Methods For Quantitative Traits Using Unrelated Populations Or Sib Pairs

Background of the Invention

The complex diseases that present the greatest challenge to modern medicine, including cancer, cardiovascular disease, and metabolic disorders, arise through the interplay of numerous genetic and environmental factors. One of the primary goals of the human genome project is to assist in the risk-assessment, prevention, detection, and treatment of these complex disorders by identifying the genetic components. Disentangling the genetic and environmental factors requires carefully designed studies. One approach is to study highly homogenous populations (Nillson and Rose 1999; Rabinow, 1999; Frank 2000). A recognized drawback of this approach, however, is that disease-associated markers or causative alleles found in an isolated population might not be relevant for a larger population. An attractive alternative is to use well-matched case-control studies of a more diverse population. A second alternative is to study siblings, inherently matched for environmental effects.

Even with a well-matched sample set, the genetic factors contributing to an aberrant phenotype may be difficult to determine. Traditional linkage analysis methods identify physical regions of DNA whose inheritance pattern correlates with the inheritance of a particular trait (Liu 1997; Sham 1997, Ott 1999). These regions may contain millions of nucleotides and tens to hundreds of genes, and identifying the causative mutation or a tightly linked marker is still a challenge. A more recent approach is to use a sufficiently dense marker set to identify causative changes directly. Single nucleotide polymorphisms, or SNPs, can provide such a marker set (Cargill et al. 1999). These are typically bi-allelic markers with linkage disequilibrium extending an estimated 10,000 to 100,000 nucleotides in heterogeneous human populations (Kruglyak 1999; Collins et al. 2000). Tens to hundreds of thousands of these closely spaced markers are required for a complete scan of the 3 billion nucleotides in the human genome. Because each SNP constitutes a separate test, the significance threshold

must be adjusted for multiple hypotheses (p-value ~ 10^{-8}) to identify statistically meaningful associations. Consequently, hundreds to thousands of individuals are required for association studies (Risch and Merikangas 1996).

The most powerful tests of association require that each individual be genotyped for every marker (Fulker et al. 1995, Kruglyak and Lander 1995, Abecasis et al. 2000, Cardon 2000) and remain far too costly for all but testing candidate genes. An alternative that circumvents the need for individual genotypes, related to previous DNA pooling methods for determination of linkage between a molecular marker and a quantitative trait locus (Darvasi and Soller 1994), is to determine allele frequencies for sub-populations pooled on the basis of a qualitative phenotype. Populations of unrelated individuals, separated into affected and unaffected pools, have greater power than related populations. If a population consists of sib-pairs, concordant pairs versus unrelated controls have greater power than discordant pairs separated into affected and unaffected pools (Risch and Teng 1998). Nevertheless, discordant designs might provide a better control for confounding factors such as age, ethnicity, or environmental effects.

The phenotypes relevant for complex disease are often quantitative, however, and converting a quantitative score to a qualitative classification represents a loss of information that can reduce the power of an association study. The location of the dividing line for affected versus unaffected classification, for example, can affect the power to detect association. Furthermore, pooling designs based on a comparison of numerical scores are not even possible with a qualitative classification scheme. These distinctions can be especially relevant when populations contain related individuals and qualitative tests have a disadvantage (Risch and Teng 1998).

There remains a need for procedures that provide phenotype associations with diseases or pathologies based on phenotypes that may be ranked on a quantitative scale. In such a scheme there is a strong need to identify procedures for optimally obtaining samples, or pooling, from a subpopulation that provide the highest assurance of displaying associations that are present. In addition there is a need to distinguish among various pooling strategies that may arise in cases with different allele frequencies and different allele correlations. There is a further need to devise a test criterion for establishing the significance of associations

between phenotypes and diseases or pathologies that may arise. The present invention addresses these and related deficiencies that currently exist.

Summary of the Invention

The present invention is based, in part, on the discovery of methods to detect an association in a population of individuals between a genetic locus and a quantitative phenotype, where two or more alleles occur at a given genetic locus, and the phenotype is expressed using a numerical phenotypic value whose range falls within a first numerical limit and a second numerical limit. These limits are used to provide for subpopulations that consist of upper and lower pools.

In some embodiments, the population of individuals includes individuals who may be classified into classes. In certain aspects of the invention, these classes are based on age, gender, race, or ethnic origin. In other aspects, some or all members of a class are included in the pools.

In various embodiments, these numerical limits are chosen so that the upper pool includes the highest 10%, 15%, 20%, 25%, 27%, 30%, or 35% of the population. In other embodiments, the numerical limits are chosen such that the lower pool includes the lowest 10%, 15%, 20%, 25%, 27%, 30%, or 35% of the population.

In one embodiment of the invention, the numerical limits are chosen to minimize false-negative errors.

In the present invention, the population of individuals can include unrelated individuals or related individuals. In one aspect, these related individuals are sibling pairs (sib pairs). In a further aspect, each member of the sib pair is selected for the upper pool. In a still further aspect, each member of the sib pair is selected for the lower pool. In still yet another aspect, neither member of the sib pair is selected. In another aspect, one member of the sib pair is selected for the upper pool and the other member of the sib pair is selected for the lower pool.

In one embodiment of the invention, sib pairs are ranked by the absolute magnitude of the difference in phenotypic value for the siblings within each pair. In one aspect, the percent of pairs with the greatest difference are identified, and the siblings in each pair are distributed such that the sibling with the high phenotypic value is selected for the upper pool and the sibling with the low phenotypic value is selected for the lower pool. In an aspect of this

embodiment, the phenotypic value of one member of the sibling pair is above a predetermined lower limit and the phenotypic value of the second member of the sibling pair is below a predetermined upper limit. In various other aspects, the percentage of pairs with the greatest difference is 80%, 70%, 60%, 54% or 50%, and the distribution provides 10%, 15%, 20%, 25%, or 27% of the population in each pool.

In an embodiment of the invention, Mahalanobis ranks are generated among sib pairs. In one aspect, these ranks are used to construct pools composed of the member of the sib pair with the more extreme Mahalanobis rank. In another aspect, the Mahalanobis ranks are used to generate a list in which the order of each member of a sib pair in this list is determined by the smaller of the distance of a member from the first member on the list and the distance of a member from the last member on the list.

Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, suitable methods and materials are described below. All publications, patent applications, patents, and other references mentioned herein are incorporated by reference in their entirety. In the case of conflict, the present specification, including definitions, will control. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

Other features and advantages of the invention will be apparent from the following detailed description and claims.

Brief Description of the Figures

Fig. 1. Shaded regions illustrate which siblings are selected under different pooling designs. The x-axis represents X_1, the phenotypic value for the first sibling, and the y-axis represents X_2, the value for the second sibling. The indicator functions I_{U1}, I_{U2}, I_{L1}, and I_{L2} take the value 1 when a sibling is selected for the denoted pool and are 0 otherwise. The unrelated-random design assumes a population of unrelated individuals, and only the first sibling is used. The pair-mean design depends on the sibling phenotype mean $X_+ = (X_1 + X_2)/2$; the pair-difference design depends on the difference $X_- = (X_1 - X_2)/2$.

Fig. 2. The population N necessary to detect association is shown as a function of the pooling fraction ρ for three values of the sibling phenotype correlation r. Panel A: $r = 0.1$, low correlation; Panel B: $r = 0.5$, moderate correlation; Panel C: $r = 0.9$, high correlation. The values of the remaining parameters are $\alpha = 5\times10^{-8}$, $1-\beta = 0.8$, $p_1 = 0.1$, $\sigma_A^2 = 0.02$, and $d/a = 0$. For low to moderate sibling correlation, the unrelated-random design is more powerful than any design using sib pairs; for high sibling correlation, sib-apart designs are more powerful. The flat minima indicate that pooling fractions close to the minima are near optimal.

Fig. 3. The population N necessary to detect association is shown as a function of the sibling phenotype correlation r. The pooling fraction ρ is optimized to minimize the population requirements at specified false-positive rate $\alpha = 5\times10^{-8}$ and power $1-\beta = 0.8$ with remaining parameters $p_1 = 0.1$, $\sigma_A^2 = 0.02$, and $d/a = 0$. Panel A: Below $r = 0.75$, the unrelated-random design is most powerful, followed by unrelated-extreme for sib pairs; above $r = 0.75$, the pair-difference design is most powerful. The sib-apart designs are more powerful than sib-together designs above $r = 0.5$ but are less powerful below this value. Panel B: The optimal pooling fraction is approximately 0.27 for the unrelated-random, pair-mean, pair-difference, and concordant designs; 0.18 for the unrelated-extreme design; and 0.23 for the discordant design. The optimal pooling fraction decreases for sib-apart designs in regions of large sibling correlation.

Fig. 4. The population N necessary to detect association is shown as a function of the minor-allele frequency p_1. The pooling fraction ρ is optimized to minimize the population requirements at specified false-positive rate $\alpha = 5\times10^{-8}$ and power $1-\beta = 0.8$ with remaining parameters $r = 0.4$, $\sigma_A^2 = 0.02$, and $d/a = 0$. Panel A: The population N is relative flat until p_1 falls below the additive variance σ_A^2, at which point the phenotype becomes nearly monogenic and the population requirement decreases. Panel B: The optimal pooling fraction ρ is relative flat until p_1 falls below the additive variance σ_A^2, at which point it decreases rapidly.

Fig. 5. The population N necessary to detect association is shown as a function of the additive variance σ_A^2. The pooling fraction ρ is optimized to minimize the population requirements at specified false-positive rate $\alpha = 5\times10^{-8}$ and power $1-\beta = 0.8$ with remaining parameters $r = 0.4$, $p_1 = 0.1$, and $d/a = 0$. Panel A: The population requirement is inversely proportional

to $1/\sigma_A^2$, except for vary large values of σ_A^2 characteristic of a monogenic trait. Panel B: The optimal pooling fraction ρ is independent of σ_A^2 except for large values of σ_A^2.

Fig. 6. The population N necessary to detect association is shown for four values of the dominance ratio d/a as a function of the pooling fraction ρ. The remaining parameters are $\alpha = 5\times10^{-8}$, $1-\beta = 0.8$, $r = 0.4$, $p_1 = 0.1$, and $\sigma_A^2 = 0.02$. Panel A: $d/a = -1$ (pure recessive); Panel B: $d/a = -0.9$; Panel C: $d/a = -0.5$; Panel D: $d/a = 1$ (pure dominant). These values were selected to sample the ratio of dominance variance to total variance for the allele, $\sigma_D^2/(\sigma_D^2 + \sigma_A^2)$. Most association methods are more sensitive to additive variance than dominance variance. Close to $d/a = 1/(2p_1 - 1)$, the additive variance vanishes and the curve of N versus ρ changes from having a shallow minimum near $\rho = 0.27$ ($\rho = 0.18$ for unrelated-extreme) to being steeply sloped toward $\rho = 0$. For rare alleles, this behavior occurs in a narrow region near $d/a = -1$ (pure recessive).

Fig. 7. The population N necessary to detect association is shown as a function of the dominance ratio d/a. Panel A: N when the pooling fraction $\rho = 0.2$; Panel B: N when ρ has been optimized to minimize the population requirements for each value of d/a; Panel C: the optimized ρ. The remaining parameters are $\alpha = 5\times10^{-8}$, $1-\beta = 0.8$, $r = 0.4$, $p_1 = 0.1$, $\sigma_A^2 = 0.02$. When $\rho = 0.2$, near-optimal for alleles with additive variance, the population requirements increase markedly near $d/a = -1$ where the additive variance is small relative to the dominance variance for a low-frequency allele. The population requirements to detect rare recessive alleles could be reduced by decreasing ρ by 10-fold to 100-fold, but this would reduce the power to detect association for alleles outside of this narrow region of large dominance variance. The population requirements and the optimal pooling fraction are not sensitive to changes in d/a for low-frequency alleles that are under-dominant ($d/a < -2$), weakly recessive ($d/a \approx -0.5$), additive ($d/a = 0$), dominant ($d/a = 1$), or over-dominant ($d/a > 1$).

Fig. 8. The population N required to detect association is shown as a function of the Type I error rate α and the Type II error rate β. The pooling fraction ρ has been optimized to minimize the population size. Panel A: N is asymptotic to 2 $\ln(1/\alpha)$ for small values of α. The remaining parameters are $1-\beta = 0.8$, $r = 0.4$, $p_1 = 0.1$, $\sigma_A^2 = 0.02$, and $d/a = 0$. Panel B: The

optimal pooling fraction ρ is not sensitive to changes in α. Panel C: The required population increases when β decreases. The remaining parameters are $\alpha = 5\times10^{-5}$, appropriate for a test of 1000 candidate polymorphisms versus a single phenotype, $r = 0.4$, $p_1 = 0.1$, $\sigma_A^2 = 0.02$, and $d/a = 0$.

Fig. 9. The repository size required to detect association using pooled DNA is shown as a function of the fraction of population ρ selected for each pool, relative to the repository size required for a regression test using individual genotyping, for a QTL making a small contribution to a complex trait. The same family structure and the same phenotypic variable, either the individual phenotype, the pair-mean, the pair-difference, or the combined results from pair-mean and pair-difference tests, are used for tests based on pooling and individual genotyping. All of these tests show the same relative efficiency as a function of pooling fraction, with an optimal fraction of 0.27 requiring only 1.24× the population for individual genotyping. The Mahalanobis design is compared to the combined regression test for a sibling phenotypic correlation of $t_R = 0.6$. The optimum occurs for this, and all other values of t_R, at ρ = 0.188.

Fig. 10. The repository size required to detect association for the Mahalanobis design, relative to the population required for a combined regression test using individual genotypes, is shown as a function of the sibling phenotypic correlation t_R.

Fig. 11. The number of individuals required for pooling designs with a sib-pair family structure is compared to the number of unrelated individuals for an association test of equivalent power and significance as a function of the sibling phenotypic correlation t_R.

Fig. 12. (A) Exact numerical results for the repository size required to detect association are shown for pooling designs as a function of σ_A^2/σ_R^2, the ratio of the additive variance of the QTL to the residual variance. The remaining parameters are allele frequency 0.1, additive inheritance, type I error 5×10^{-8}, and type II error 0.2. (B) The allele frequency difference at significance is shown for the same parameters as in Fig. 12A. In this an all subsequent figures, unrelated-population is a dotted line, Mahalanobis a thin line, pair-mean a dashed line, pair-difference a dot-dashed line, and sib-combined a thick line.

and 0 for A_2A_2. The bivariate probability distribution $P(G_1,G_2)$ of the 9 possible combinations of dizygotic sib-pair genotypes G_1 and G_2, shown in Table I, can be derived by considering all possible parental mating types and their offspring genotype distributions (Neale and Cardon 1992). The shared genetic makeup implies that $P(G_1,G_2) \neq P(G_1)P(G_2)$.

Using the notation defined above, the effect μ_G of genotype G on the phenotype is $a-\mu$, $d-\mu$, and $-a-\mu$ for genotypes A_1A_1, A_1A_2, and A_2A_2 respectively. The constant $\mu = a(p_1 - p_2) + 2dp_1p_2$ ensures that the phenotype has zero mean. The ratio d/a, termed the dominance ratio, is –1 for a recessive allele, +1 for a dominant allele, and 0 for an additive allele.

The phenotypic variance contributed by the genotype G can be partitioned into an additive component σ_A^2 and an dominance component σ_D^2, with

$\sigma_A^2 = 2pq[a-d(p-q)]^2$, and

$\sigma_D^2 = 4p^2q^2d^2$.

In a population of unrelated individuals, the distribution $f[X]$ of trait values is a mixture of 3 univariate normals, one for each genotype:

$f[X] = \Sigma_G f[X|\mu_G]P(G)$, with

$f[X|\mu_G] = (2\pi\sigma_R^2)^{-1/2}\exp[-(X-\mu_G)^2/2\sigma_R^2]$

and the residual variance $\sigma_R^2 = 1-\sigma_A^2-\sigma_D^2$.

Similarly, in a population of sib pairs, the bivariate distribution of trait values $f[X_1X_2]$ is a mixture of 9 bivariate normals, appropriately weighted according to genotype combination:

$$f[X_1,X_2] = \sum_{G_1G_2} f[X_1,X_2|G_1,G_2]P[G_1,G_2].$$

The mean of X_j is μ_{G_j} for sib $j = 1$ or 2; both X_1 and X_2 have residual variance $\sigma_R^2 = 1 - \sigma_A^2 - \sigma_D^2$; and X_1 and X_2 have correlation r_R due to shared residual polygenic effects and environmental factors. The total correlation r between sib pairs, including effects from genotype G, is

$r = r_R + \sigma_A^2/2 + \sigma_D^2/4$.

It is convenient to re-express the phenotypes of sib pairs in terms of X_+ and X_-, defined as the linear combinations $X_\pm = (X_1 \pm X_2)/2$, because these components are uncorrelated and the

probability distribution $f[X_+,X_-|G_1,G_2]$ factors into the product $f[X_+|G_1,G_2]\,f[X_-|G_1,G_2]$. The individual probability distributions for X_+ and X_- are

$f[X_\pm|G_1,G_2] = (2\pi\sigma_\pm^2)^{-1/2}\exp[-(X_\pm-\mu_\pm)^2/2\sigma_\pm^2]$, with

$\mu_\pm(G_1,G_2) = (\mu_{G_1} \pm \mu_{G_2})/2$ and

$\sigma_\pm^2 = \sigma_R^2(1 \pm r_R)/2$.

Allele frequencies $p_\pm$ are similarly defined as $p_\pm(G_1,G_2) = (p_{G_1} \pm p_{G_2})/2$.

2.2 Test Statistic and the Null Hypothesis

We consider tests in which an upper and lower pool, each containing n individuals, are selected according to higher and lower phenotypic values from a larger population of N individuals. The frequencies p_U and p_L of allele A_1 are calculated for the upper and lower pools, and the frequency difference is converted to the test statistic T,

$T = (p_U - p_L)/(\sigma_0/\sqrt{n})$.

The variance $p_U - p_L$ under the null hypothesis that genotype G has no effect on phenotype X is $\mathrm{Var}(p_U - p_L) = \sigma_0^2/n$. When the null hypothesis is valid and n is large, T follows a standard normal distribution and σ_0 is independent of n.

The value of σ_0 depends on the population allele frequencies and also on the method used to select the n individuals for each pool. Specifically, let n_C be the total number of sib pairs selected for the same pool and n_D be the number split between pools, with the remaining $2(n - n_C - n_D)$ individuals unrelated. The contribution of the unrelated individuals to $\mathrm{Var}(p_U-p_L)$ is $[2(n-n_C-n_D)/n^2]\mathrm{Var}(p_G)$, and the individual variance is

$\mathrm{Var}(p_G) = p_1^2(1) + 2p_1p_2(1/4) - p_1^2 = p_1p_2/2$.

The contribution of the pooled-together sib-pairs is

$[n_C/n^2]\mathrm{Var}(p_{G_1} + p_{G_2}) = [n_C/n^2][2\mathrm{Var}(p_G) + 2\mathrm{Cov}(p_{G_1}, p_{G_2})] = (n_C/n^2)(3p_1p_2/2)$

because the covariance between genotypes in a sib-pair is half the individual variance, reflecting that sibs share half their genetic material. Similarly, the contribution of the pooled-apart sib-pairs is

$[n_D/n^2]\mathrm{Var}(p_{G_1} - p_{G_2}) = [n_D/n^2][2\mathrm{Var}(p_G) - 2\mathrm{Cov}(p_{G_1}, p_{G_2})]$.

The result for σ_0^2 is

$\sigma_0^2 = [1 + (n_C/2n) - (n_D/2n)]p_1p_2$,

with important limiting cases of $p_1p_2/2$ for pure sib-apart pooling, p_1p_2 for pure unrelated pooling, and $3p_1p_2/2$ for pure sib-together pooling.

The allele frequency p_1 may be determined from the entire population. It is also possible to estimate p_1 as the mean $(p_U + p_L)/2$, which is closer to 0.5 than the population mean p_1 in the case of true association. The resulting σ_0 is larger, and using the mean results in a conservative test.

2.3 Pooling Design

A pooling design is a set of rules to determine which sibs are selected for the upper and lower pools. For an unrelated population, these rules take the form of a pair of indicator functions $I_U(X)$ for the upper pool and $I_L(X)$ for the lower pool. Each function takes the value 1 if an individual is selected for the specified pool and is 0 otherwise. In general, individuals are selected for at most one pool and $I_U + I_L$ is either 0 or 1.

The rules for sib-pairs may be formulated in terms of four indicator functions which depend on both sibling phenotypic values X_1 and X_2. These indicator functions may be written $I_{Sj}(X_1,X_2)$ or equivalently $I_{Sj}(X_+,X_-)$, where the side S is U or L and $j = 1$ or 2 labels the sibling. The indicator function has value 1 if sib j is selected for side S and is 0 otherwise. As before, each individual is selected for at most one pool and $I_{Uj} + I_{Lj}$ is either 0 or 1.

A summary of pooling designs in terms of the indicator functions is provided in Table II. The indicator functions are specified by upper and lower phenotype thresholds X_U and X_L and the Heaviside step function H(x),

$$H(x) = \begin{cases} 1, & x > 0; \\ 1/2, & x = 0; \\ 0 & x < 0. \end{cases}$$

The values of X_U and X_L are defined implicitly by the requirement that the upper pool and lower pool each contains a fraction ρ of the total population.

Three types of designs are considered: unrelated pooling designs, in which none of the $2n$ pooled individuals are related (although the individuals may be drawn from a larger population of related individuals); sib-together pooling designs, in which each pool consists of $n/2$ sib pairs; and sib-apart poolingdesigns, in which n sib pairs are split between the upper and lower pools.

Unrelated Pooling Designs

Two types of unrelated pools are shown. The first, unrelated-random, pools the n individuals with the highest and lowest phenotypic values from a population of N unrelated individuals. The term random arises because the N unrelated individuals may be obtained by selecting one sib at random from an initial population of N sib pairs.

The second unrelated design, unrelated-extreme, first reduces a population of $N/2$ sib pairs to $N/2$ unrelated individuals by selecting the individual with the more extreme phenotypic value from each sib pair. Tails with n individuals are then selected for pooling from this unrelated sub-population. The more extreme sib is defined as having a greater distance $|X_i|$ from the phenotype mean. Other definitions of distance, such as the distance from the phenotype median, or non-parametric definitions, such as the phenotype percentile score, are also possible and yield similar results for a normal distribution of phenotype scores.

Sib-Together Pooling Designs

Two sib-together designs are analyzed, each starting with a population of N individuals in $N/2$ sib pairs. The first, termed concordant, is analogous to concordant pooling based on a qualitative, affected/unaffected classification. If both sibs have phenotypic values above an upper threshold X_U, the pair is selected for the upper pool; if both values are below a lower threshold X_L, the pair is selected for the lower pool. The thresholds are adjusted until $n/2$ pairs have been added to each pool. The second sib-together design, pair-mean, is based on the phenotype mean X_+ for each pair: above X_U and the pair is selected for the upper pool; below X_L and the pair is selected for the lower pool.

Sib-Apart Pooling Designs

Two sib-apart designs are also analyzed, each starting with $N/2$ sib pairs. The first is termed discordant, again analogous to qualitative discordant pooling. If one sib in a pair has a

phenotypic value above an upper threshold X_U and the other has a value below a lower threshold X_L, the sib with the higher value is selected for the upper pool and the sib with the lower value is selected for the lower pool. The thresholds X_U and X_L must have an additional constraint in order to arrive at a unique solution. The constraint used here is that the thresholds straddle the phenoype mean and are equidistant from it. Other constraints, such as at equal percentiles away from the median phenotype, are possible but give similar results for a normal distribution of phenotype scores.

The second sib-apart design, termed pair-difference, selects the n sib pairs with the greatest magnitude of difference $|X_1-X_2|$ in phenotypic values. The sib with the higher value is selected for the upper pool and the sib with the lower value enters the lower pool. Again, more general measures of distance are possible.

The depiction of pooling designs in Fig. 1 complements the mathematical description. Each of the six panels displays one of the pooling designs identified above. The coordinate axes are X_1 and X_2, the sib-pair phenotypic values, and cross at the overall phenotype mean of 0. Areas in the graph are shaded when one or more of the indicator functions is 1. In the unrelated-random design at the upper left, for example, an unrelated population is generated by taking the first sib from each pair and the pooled regions are vertical half-planes. If the second sib had been taken from each pair, the half-planes would be horizontal. The panel in the upper right depicts the unrelated-extreme pools. The regions corresponding to sib 1 being extreme are the two triangles bordered by $X_1 = \pm X_2$ and along the horizontal axis. These regions are truncated at the upper threshold X_U and the lower threshold X_L to yield the contribution of sib 1 to the upper and lower pools. Sib 2 makes similar contributions, symmetric across the $X_1 = X_2$ axis. This panel shows an example where $X_U \neq -X_L$, which is the general case when the phenotype mean and median do not coincide. When equality holds, the excluded region in the center is perfectly square.

The middle panels depict the two sib-together designs. On the left is the concordant design: to be selected for pooling, both sibs must be above or below a threshold. The upper threshold X_U could also provide the definition for a qualitative classification affected/unaffected. In this case, the vertex of the lower pool moves northeast to meet the vertex of the upper pool at the

phenotypic values X_U,X_U. The panel to the right shows the pair-mean design. Here, sib pairs are selected if their mean X_+ exceeds an upper threshold X_U or falls below a lower threshold X_L. The orthogonal coordinate X_- is uncorrelated with X_+ and unconstrained in this design. Note that the boundary lines $X_+ = X_U$ and $X_+ = X_L$ have intercepts $2X_U$ and $2X_L$ in the X_1-X_2 coordinate system.

The bottom panels depict the discordant design on the left and the pair-difference design on the right. The discordant design selects sib-pairs from rectangular regions in the upper left and lower right; the pooling boundaries in the pair-difference design are lines of constant X_-, with X_+ unconstrained.

Despite the close analogy, there is an important difference between the concordant and discordant designs described here for quantitative traits and the designs described elsewhere for qualitative traits (Risch and Teng, 1998). In this formulation for quantitative traits, the upper and lower thresholds define tails of a population distribution and a sizeable population fraction falls between the tails. In a typical formulation for qualitative traits, and especially for qualitative traits without an obvious quantitative basis, a single threshold divides the population into two classes: a smaller affected class and a larger unaffected class holding most of the population. In the terminology used here, such designs have $X_U = X_L$.

2.4 Distribution of p_U–p_L under the alternative hypothesis

The fraction ρ_S of the total population selected for each pool may be written

$$\rho_S = \sum_{G_1,G_2} \sum_{j=1,2} \rho_{Sj}(G_1,G_2),$$

where, as before, S = U or L labels the upper or lower pool and

$$\rho_{Sj}(G_1,G_2) = (1/2)\, P(G_1,G_2) \int_{-\infty}^{\infty} dX_+ \int_{-\infty}^{\infty} dX_- f[X_+|\, G_1,G_2] f[X_-|\, G_1,G_2]\, I_{Sj}(X_+,X_-)\,.$$

The initial factor of (1/2) arises because the phenotype and genotype distributions are normalized to 1 per sib-pair rather than 2. In practice, the upper and lower thresholds X_U and X_L are adjusted until the fraction in each pool is $\rho < 1$. For an unrelated population or for a sib-pair population pooled with the pair-mean or pair-difference design, the largest possible ρ is 0.5 and the entire population splits evenly into two pools. The concordant and discordant

designs have a maximum ρ that is smaller than 0.5 because, as can be seen from Fig. 1, these designs always exclude quadrants of the total population. For a sib-pair population with the unrelated-extreme design, the largest possible ρ is 0.25.

For feasible values of ρ, the expected allele frequency in pool S is

$$p_S = \rho^{-1} \sum_{G_1,G_2,j} \rho_{Sj}(G_1,G_2)\, p_{G_j},$$

where p_{G_j} is the allele frequency of the j^{th} sib of the pair and the expected number of such sibs selected for the pool is $n\rho^{-1}\rho_{Sj}(G_1,G_2)$. These numbers follow a multinomial distribution, with the following general properties: when a random variable $x = n^{-1}\Sigma_i n_i x_i$ with the index i ranging over a discrete set of sub-populations, the total number of samples $n = \Sigma_i n_i$ fixed, x_i fixed for all samples from sub-population i, the expectation values $n_i/n = \theta_i$ fixed, and $\Sigma_i\theta_i = 1$, then the expectation value of x is $\Sigma_i\theta_i x_i$ and its variance $\mathrm{Var}(x) = n^{-1}\{\Sigma_i\theta_i x_i^2 - (\Sigma_i\theta_i x_i)^2\}$ (Beyer, 1984). Using these results for a multinomial distribution, the variance of the test statistic under the alternative hypothesis is written

$$\mathrm{Var}(p_U - p_L) = \sigma_1^2/n$$

where σ_1^2 is independent of the number of individuals n per pool.

For the unrelated-extreme design, p_U and p_L are independent multinomial distributions and

$$\sigma_1^2 = \rho^{-1} \sum_{G_1,G_2,j} \{\rho_{Uj}(G_1,G_2)\,(p_{G_j}^2 - p_U^2) + \rho_{Lj}(G_1,G_2)(p_{G_j}^2 - p_L^2)\}.$$

For the unrelated-random design, the index j is irrelevant, yielding simpler expressions:

$p_S = \rho^{-1}\Sigma_G \rho_S(G) p_G$, and

$\sigma_1^2 = \rho^{-1}\Sigma_G \{\rho_U(G)(p_G^2 - p_U^2) + \rho_L(G)(p_G^2 - p_U^2)\}$.

For the sib-together designs, $I_{S1} = I_{S2}$ and the expected allele frequencies are

$$p_S = \rho^{-1} \sum_{G_1,G_2} 2\rho_{S1}(G_1,G_2) p_+.$$

The corresponding frequencies θ_i for the multinomial distribution are $2\rho^{-1}\rho_{S1}(G_1,G_2)$ and the effective number of samples is $n/2$. The resulting variance term is

$$\sigma_1^2 = 2\rho^{-1} \sum_{G_1,G_2} \{2\rho_{U1}(G_1,G_2)(p_+^2 - p_U^2) + 2\rho_{L1}(G_1,G_2)(p_+^2 - p_L^2)\}.$$

For the sib-apart designs, $I_{U1}=I_{L2}$ and $I_{L1}=I_{U2}$. The expectation value of the allele frequency difference is

$$p_U - p_L = \rho^{-1} \sum_{G_1,G_2} \rho_{U1}\, p_{G_1} + \rho_{U2}\, p_{G_2} - \rho_{L1}\, p_{G_1} - \rho_{L2}\, p_{G_2} = \rho^{-1} \sum_{G_1,G_2} 2\rho_{U1} p_- + \rho^{-1} \sum_{G_1,G_2} 2\rho_{L1}(-p_-).$$

Due to the symmetry between the two siblings, $\rho^{-1}\Sigma_{G_1,G_2} 2\rho_{U1} = \rho^{-1}\Sigma_{G_1,G_2} 2\rho_{L1} = 1$, and $p_U - p_L$ is the sum of two multinomial distributions each with expectation value $(p_U-p_L)/2$. The effective number of samples for each distribution is $n/2$, and the variance term is

$$\sigma_1^2 = 2\rho^{-1} \sum_{G_1,G_2} 2(\rho_{U1} + \rho_{L1})\{p_-^2 - [(p_U-p_L)/2]^2\}.$$

When the null hypothesis is valid, each of these expressions for σ_1 reduces to the corresponding expression for σ_0. If the alternative hypothesis is valid, σ_1 is smaller than σ_0 to the extent that variance in the test statistic is explained by the pooling design. Nevertheless, in most cases σ_0 is an excellent approximation.

2.5 Power

The statistical power $1-\beta$ to reject the null hypothesis for a single one-tailed test with p-value α, where α is equivalent to the false-positive rate or Type I error rate and β is equivalent to the false-negative rate or Type II error rate, is

$$1-\beta = 1 - \Phi\{ [z_\alpha \sigma_0 - \sqrt{n}\,(p_U - p_L)] / \sigma_1 \},$$

where $\Phi(z)$ is the cumulative standard normal distribution, $1-\Phi(z_\alpha) = \alpha$. Solving for n and using the relation $n/N = \rho$, the total number of individuals N necessary to generate pools with the required power is

$$N = \rho^{-1} [(z_\alpha \sigma_0 - z_{1-\beta}\sigma_1) / (p_U - p_L)]^2,$$

where $\rho = n/N$ is the fraction of the total population selected for each pool. In either case, replacing σ_1 with σ_0 would result in a conservative test.

2.6 Computational Methods

Exact results for the distribution of the test statistic T under the null hypothesis and under the alternative hypothesis, subject only to the approximation that T is normal, were obtained by numerical computations converged to better than 1 part in 10^6 (Press et al. 1997). Brent's root-

finding algorithm was used to determine the threshold values X_U and X_L for the upper and lower pools for a given pooling design and pooling fraction ρ; Brent's optimization algorithm was then used to find the ρ with maximum power. The integrals providing $p_U - p_L$ and σ_1^2 were evaluated numerically using Romberg integration with a change of variables to the reciprocal for infinite integration limits. Integration was restricted to regions where an indicator function was non-zero. In order to reduce computational requirements, the final integral of a normal distribution over fixed limits was evaluated using a polynomial approximation to the error function. This technique reduced the two-dimensional integrals over bivariate normals to one-dimensional integrals for the unrelated-extreme, concordant, and discordant designs, while integration was avoided completely for the unrelated-random, pair-mean, and pair-difference designs. The 9 sib-pair genotypes were reduced by symmetry to 5 genotypes for further savings. Using a 750 MHz Pentium III running Linux, the root-finding and minimization for each parameter set required less than 0.01 sec each for the unrelated-random, pair-mean, and pair-difference designs and approximatley 6 sec each for the unrelated-extreme, concordant, and discordant designs.

The numerical results, and the underlying theory, are robust when n, the number of individuals per pool, is large and $2(p_U + p_L)n$, the number of alleles in the pools, approximately follows a normal distribution. In certain regions of extreme parameter values, however, the numerical solution for n drops below 1. This behavior signals a breakdown of various assumptions of the theory, and results in these regions are unreliable.

The properties and characteristics of the methods of the present invention are set forth in the Examples. It is shown, for example, that the optimal design for unrelated individuals is to pool the top and bottom 27% of the population. This design using N unrelated individuals has greater power than designs using $N/2$ sib pairs when the phenotypic correlation between sibs is low to moderate, below 75%, but has less power than sib pair designs when the correlation is above 75%.

Of the designs explored for a population of sib pairs, the unrelated-extreme design is the best for low to moderate sibling phenotype correlation. In this design, the more extreme sib is selected from each pair, then the top and bottom 36% of this subset are pooled. When the correlation is high, above 75%, the best design found for sib pairs is to first select the 27% of

pairs with the greatest phenotype difference, then split each pair by phenotypic value to form an upper and lower pool. The pair-difference design might also be applied at low to moderate sibling correlation to reduce the rate of spurious association due to population stratification. The optimal pooling fractions for these designs were determined by minimizing the population requirements. The minima were generally quite flat, and pooling fractions close to the optimal fractions give near-optimal results.

Compared with the results obtained by others for pooling based on qualitative traits, the results derived using the methods of the present invention for quantitative traits are thought to be surprising. For earlier pooling strategies based on qualitative traits, designs using unrelated individuals were found to be more powerful than designs using sib pairs; when populations were restricted to sib pairs, concordant designs were found to have greater power than discordant designs (Risch and Teng 1998). In contrast, for quantitative phenotypes, the methods of the present invention indicate that unrelated individuals become less powerful than sib pairs when sibling correlation is high, and that sib-apart designs become more powerful than sib-together designs when the sibling correlation is above 50%. This result is significant because highly heritable traits that are likely to be the first targets of large-scale genotyping studies often exhibit sibling correlations of 50% or higher. Quantitative phenotypic values also permit the use of the unrelated-extreme design, which does not have an obvious analog for qualitative phenotypes that categorize individuals as affected/unaffected.

The sib-together and sib-apart pooling designs of the present invention, which draw individuals from extreme-high and extreme-low phenotypes, are anticipated to be more powerful than alternative designs that compare one extreme to the remainder of the population, as in a qualitative affected/unaffected classification. The affected/unaffected classification establishes a single threshold for a quantitative phenotype, and the allele frequency in the large unaffected class is close to the population mean. In contrast, the quantitative designs of the present invention employ two thresholds, and the allele frequencies in the upper and lower pools are approximately equidistant from the population mean. The allele frequency difference between pools is consequently half as large for the qualitative design as for the quantitative design of the present invention, and the population requirements are four times as large, or half as large if the overall allele frequency is assumed to be known exactly. These conclusions are similar to those reached in the context of linkage analysis for

quantitative trait localization using extremely concordant and extremely discordant sib pairs (Risch and Zhang 1995, Risch and Zhang 1996, Zhang and Risch 1996, Gu et al. 1996).

As with most genotyping designs, the pooling strategies described here are primarily sensitive to the additive variance from an allele. Since the additive variance for an allele is approximately equal to the fraction of heterozygotes times the square of half the phenotype shift between the two homozygotes, rare alleles with larger phenotype shifts may be detected with the same power as common alleles with smaller shifts. When the allele frequency becomes smaller than the additive variance of the allele, however, the frequency shift must become very large to compensate and the phenotype begins to resemble a monogenic trait.

The results provided here also imply the precision required for allele frequency determinations for pooled DNA. Approximately 3000 individuals are required for a genome-wide screen with an optimal pool size n of 600 to 800 individuals. The frequency difference corresponding to significance at $\alpha=5\times10^{-8}$ ($z_\alpha = 5.33$) for a polymorphism with minor-allele frequency p_1 is $z_\alpha[p_1(1-p_1)/n]^{1/2}$, which is 5% for an allele frequency of 0.1 and 2% for an allele frequency of 0.01. An experimental measurement should provide an order of magnitude better precision in the allele frequency difference to avoid losing information.

3. Examples for Mod I 1

Overview to the Examples

In this section, total population sizes are presented for a wide range of parameters and as functions of the pooling fraction ρ. The first parameters explored are the sib-pair phenotype correlation r and the allele frequency p_1; these parameters are readily determined experimentally at the start of an association study. The next set of parameters explored are the additive phenotype variance σ_A^2, the dominance ratio d/a, and the resulting dominances variance σ_D^2 and genotype effects μ_G, which are not known at the start of a study. Finally, the dependence of the population requirements on the false-positive rate α and false-negative rate β is explored. As each single parameter is varied in turn, the remaining parameters are held fixed at a set of values selected to serve as a common reference.

The reference value for sibling phenotype correlation was based on reported values for genetic heritabilities and shared environmental factors. Estimates of the genetic heritability for complex traits range from 20% for cancer (Verkasalo et al. 1999), 20% to 40% for Type 2 diabetes mellitus (NIDDM) (Watanabe et al. 1999), 50% for pulmonary function (Wilk et al. 2000), 10% to 50% for systolic and diastolic blood pressure (Iselius et al. 1983, Perusse 1989), and 70% to 90% for cholesterol level (Austin et al. 1987). Shared environmental factors are estimated to contribute 7% of the overall phenotype variance for cancer (Verkasalo et al. 1999), 20% to 40% for blood pressure (Iselius et al. 1983, Perusse et al. 1989), and 15% for serum lipid levels (Heller et al. 1993). The sibling phenotype correlation, equal to half the genetic heritability plus the shared environmental contribution, varies over a wide range for these traits. A phenotype correlation of 40%, in the middle of the range, was selected to serve as the reference.

Reported minor-allele frequencies for SNPs found in multiple populations range from 5% to 25%, with lower frequencies for variations which cause non-conservative amino acid changes and higher frequencies for conservative substitutions and changes in non-coding regions (Cargill et al. 1999, Goddard et al. 2000). A reference value of 10% was selected for p_1, typical of changes in the coding region.

The genetic variance arising from a typical SNP was modeled by assuming that the genetic heritability arises from multiple loci, each of which makes an independent contribution with a characteristic size equal to the genetic heritability divided by the total number of contributing loci. Assuming that approximately 20 polymorphic sites contribute to a genetic heritability of 40% yields a reference value of 0.02 for $\sigma_A^2 + \sigma_D^2$. The reference value selected for the dominance ratio was $d/a = 0$, indicating a purely additive allele.

In practice, the false-positive rate α is matched to the number of individual tests that are to be conducted in an association study. For a genome scan of 10^6 individual markers versus a single phenotype, for example, or for a scan of 10^4 markers versus 100 distinct phenotypes, a false-positive rate α per marker should be no more than 5×10^{-8} for a final p-value < 0.05 for the detection of an association. If only 1000 markers are used, for example as in a test of candidate polymorphisms, then the value $\alpha = 5\times10^{-5}$ suffices. The false-positive rate selected as a reference was $\alpha = 5\times10^{-8}$ ($z_\alpha = 5.33$), a value suggested to provide a sufficiently low number of false positives after applying a multiple-hypothesis-testing correction corresponding to a full-genome scan (Risch and Merikangas 1996). The power $1-\beta$ was fixed at 0.8 ($z_{1-\beta} = -0.84$) for a 20% false-negative rate.

Figures depicting the results use a consistent scheme. The unrelated designs are represented as solid lines, thin for unrelated-random and thick for unrelated-extreme; the sib-together designs are represented as equal-spaced dashed lines, thin for concordant and thick for pair-mean; and the sib-apart designs are represented as unequally-spaced dashed lines, thin for discordant and thick for pair-difference.

Example 1. Sibling Phenotype Correlation

The minimum population size N required to detect association as a function of the sibling phenotype correlation r and the pooled fraction ρ is shown in Fig. 2, with the remaining parameters at their previously defined reference values ($\alpha = 5\times10^{-8}$, $1-\beta = 0.8$, $p_1 = 0.1$, $\sigma_A^2 = 0.02$, $d/a = 0$). The three panels in Fig. 2 show a range of sibling phenotype correlations: $r = 0.1$ (Panel A), 0.5 (Panel B), and 0.9 (Panel C). In each panel, as the pooling

fraction increases from $\rho = 0$, each design has a sharp then more gradual decrease in population requirements. Eventually N attains a minimum, indicating the optimal pooling fraction for maximum power, and then gradually increases with ρ. A second feature seen in all three panels is the similarity between the unrelated designs, between the sib-together designs, with pair-mean always more powerful than concordant, and between the sib-apart designs, with pair-difference always more powerful than discordant. Furthermore, for larger values of ρ the required numbers of concordant and discordant sib pairs are not met.

In Fig. 2, Panel A shows that for small values of the phenotype correlation the design with the greatest power is unrelated-random, with unrelated-extreme slightly less powerful. The sib-together designs require approximately twice as large a sample, and the sib-apart designs require three to four times as many. In Panel B, at the intermediate phenotype correlation $r = 0.5$, the unrelated designs are still the most powerful, while the sib-together designs have increased population requirements and the sib-apart designs have decreased to meet in the middle. At large values of the sibling phenotype correlation, $r = 0.9$ in Panel C, the sib-apart designs are most powerful. The unrelated designs require approximately twice as large a population, and the sib-together designs have far greater requirements.

The regions near the minima of N for each design are quite flat, indicating that pooling fractions within 0.1 of the minimum may give near-optimal results. The exact values of these minima are depicted in Fig. 3. The population requirements are shown in Panel A, and the corresponding optimal pooling fractions are shown in Panel B. The unrelated-random design is insensitive to the sibling correlation r, as seen in Panel A, as is the unrelated-extreme design except at the highest values of r. The sib-together designs require larger populations as r increases, while the sib-apart designs require smaller populations. The sib-together and sib-apart designs cross near $r = 0.5$, and the sib-apart and unrelated designs cross near $r = 0.75$. The optimal pooling fractions are insensitive to the changes in the sibling correlation for values below $r = 0.75$, as seen in Panel B. The unrelated-random, pair-mean, pair-difference, and concordant designs have an optimal ρ near 0.27 in this region of low to moderate correlation, while the unrelated-extreme design has an optimum near $\rho = 0.18$ and the discordant design near 0.23. For phenotypes with high correlation, $r > 0.75$, the optimal fraction for ρ decreases and only highly discordant sibs are selected for the sib-apart designs.

Example 2. Allele Frequency

The results of changing the allele frequency p_1 while optimizing the pooling fraction and holding the remaining parameters constant at their reference values ($\alpha = 5\times10^{-8}$, $1-\beta = 0.8$, $r = 0.4$, $\sigma_A^2 = 0.02$, $d/a = 0$) are shown in Fig. 4. The population requirements corresponding to the optimal pooling fraction ρ are shown in Panel A, and the corresponding fractions ρ in Panel B. The dependence on p_1 is symmetric about p_1=0.5; results are shown only for the region $p_1 < 0.5$ and are displayed on a logarithmic scale to highlight the behavior at low allele frequency.

At moderate frequencies of the minor allele, $p_1 > 1\%$, the power and pooling fraction are both insensitive to the allele frequency. This behavior, which arises when σ_A^2 is held constant and changes in μ_G are allowed to compensate for changes in p_1, is often observed in variance components models (Liu 1997) . Thus, as long as the allele frequency is not too small, lower frequency alleles with larger effects and higher frequency alleles with smaller effects are found with similar power.

At smaller allele frequencies, $p_1 < 1\%$, the increasingly rare allele has an corresponding large effect μ_G on the phenotype, and the population requirements decrease. The crossover into this region occurs when the allele frequency p_1 falls below its contribution $\sigma_A^2 + \sigma_D^2$ to the overall phenotypic variance. The pooling fraction also decreases with p_1 in this region. The exception to this trend is the discordant design, which has a dramatic drop in power for low frequency alleles.

Example 3. Additive Allele Variance

The population size N required to detect association is shown as Panel A in Fig. 5 as a function of the additive phenotypic variance arising from genotype G, with the remaining parameters fixed at their reference values ($\alpha = 5\times10^{-8}$, $1-\beta = 0.8$, $r = 0.4$, $p_1 = 0.1$, $d/a = 0$). The population size and the variance have a clear inverse linear relationship over three orders of

magnitude. This behavior corresponds to $N \propto (p_U - p_L)^{-2}$ with p_U and p_L proportional in turn to σ_A.

The corresponding optimal pooling fractions are shown in Fig. 5, Panel B. Over most of the range, $\sigma_A^2 < 0.1$, the optimal fractions are not sensitive to the variance arising from the allele. At larger values of the variance the phenotype becomes nearly monogenic and smaller pooling fractions and populations are required.

Example 4. Recessive, Additive, and Dominant Alleles

The series of panels in Fig. 6 depicts the required population size as a function of the pooling fraction ρ for a range of dominance ratios d/a. The values for d/a were selected to provide adequate sampling of the ratio of the dominance variance to the additive variance. This contribution, $\sigma_D^2/(\sigma_A^2 + \sigma_D^2)$, is 82% at $d/a = -1$ (pure recessive), 65% at –0.9, 11% at –0.5, and 5% at +1 (pure dominant). The remaining parameters were held at their reference values ($\alpha = 5\times10^{-8}$, $1 - \beta = 0.8$, $r = 0.4$, $p_1 = 0.1$, $\sigma_A^2 = 0.02$, $d/a = 0$). The pooling fraction was set to $\rho = 0.2$ for this series of panels and represents a near-optimal fraction for additive variance, $d/a = 0$.

For pure recessive traits, $d/a = -1$ in Panel A (82% dominance variance for $p_1 = 0.1$), the estimate for N approaches an apparent minimum at $\rho = 0$, and the assumption of normality of the test statistic is no longer valid. When d/a is –0.9 and the dominance variance contribution has dropped to 65%, the curves for N in Panel B start to flatten, and when $d/a = -0.5$, in which the heterozygote mean is still three-quarters of the way towards the minor-allele homozygote, the curves in Panel C are nearly indistinguishable from the results for pure additive (not depicted) and pure dominant, Panel D.

These results again signal that pooling methods for quantitative phenotypes are more sensitive to changing additive variance than to changing dominance variance. The dominance variance is only significant in regions where the additive variance vanishes, $d/a = 1/(p_1 - p_2)$. This

region occurs near –1 for a low-frequency allele, indicating that association studies have weak power to detect low-frequency recessive alleles or their high-frequency dominant counterparts.

These effects are shown in greater detail in Fig. 7. The population requirements are shown as a function of the dominance ratio d/a at the fixed pooling fraction $\rho = 0.2$ in Panel A and at the optimal fraction in Panel B. Other than the region in which the additive variance vanishes, $d/a = -1.125$ for $p_1 = 0.1$, the results in both panels are similar and show little dependence on d/a. This is true even in regions of strong over-dominance, $d/a > 1$, and under-dominance, $d/a < -2$. Near the region of vanishing additive variance the optimal pooling fraction ρ drops rapidly, as seen in Panel C, and the results for the optimal ρ and $\rho = 0.2$ differ.

Example 5. False-Positive Rate and False-Negative Rate

When the widths of the distribution of the test statistic under the null and alternative hypothesis are approximately equal, the equation for the population necessary to detect association has the form $N \propto (z_\alpha - z_{1-\beta})^2$. When α becomes small, the behavior $z_\alpha \sim [-2 \ln(\alpha)]^{1/2}$ for small α, extracted from an asymptotic expansion for $\Phi(z)$ (Mathews and Walker 1970), leads to the asymptotic behavior $N \sim 2 \ln(1/\alpha)$, which is seen clearly as the linear behavior in Panel A of Fig. 8. The remaining parameters are fixed at their reference values ($1 - \beta = 0.8$, $r = 0.4$, $p_1 = 0.1$, $\sigma_A^2 = 0.02$, $d/a = 0$). Compared to a whole-genome scan with $\alpha = 5\times10^{-8}$ ($z_\alpha = 5.33$) and a 20% false-negative rate, for example, which requires 2400 individuals pooled with the unrelated-random design or 3000 siblings pooled with the unrelated-extreme design, a test of 1000 candidate polymorphisms with $\alpha = 5\times10^{-5}$ ($z_\alpha = 3.89$) requires 1400 unrelated individuals or 1800 siblings, while a test for association between a single polymorphism and a single phenotype, $\alpha = 0.05$ ($z_\alpha = 1.64$), require 400 unrelated individuals or 500 siblings. The optimal fraction ρ for pooling is not sensitive to the choice for α itself, as seen in Panel B.

The effects of varying the false-negative rate β are similar to the effects of varying α because the population requirements depend predominantly on the difference $z_\alpha - z_{1-\beta}$ rather than on the value of either alone. For small values of β, $N \sim 2 \ln(1/\beta)$. This linear behavior is demonstrated in Panel C, where the remaining parameters have their reference values except

for $\alpha = 5\times10^{-5}$ corresponding to a test of candidate polymorphisms. The optimal pooling fraction p does not depend sensitively on β, as shown in Panel D.

4. Model 2

4.1 Variance components model

A standard variance components model is used to describe the joint phenotype-genotype probability distribution. A quantitative phenotype X, standardized to mean 0 and variance 1, is hypothesized to be affected by the genotype G at a biallelic locus with minor allele A_1 and major allele A_2 occurring at population frequencies p and $1-p$. More generally, A_2 may represent any of a number of alternate alleles, and $1-p$ their aggregate frequency. The population is assumed to be random mating and in Hardy-Weinberg equilibrium. The symbol P is used to denote a probability, and the genotype frequencies $P(G)$ are p^2, $2p(1-p)$, and $(1-p)^2$ for A_1A_1, A_1A_2, and A_2A_2 respectively. The frequency of allele A_1 in genotype G, denoted p_G, is 1 for A_1A_1, 0.5 for A_1A_2, and 0 for A_2A_2. The variance of the allele frequency for an individual, denoted σ_p^2, is $p(1-p)/2$.

The frequency of a genotype combination for a sib pair is denoted $P(G_1,G_2)$. Only full sibs are considered. The probability distribution $P(G_1,G_2)$ of the 9 possible combinations of sib-pair genotypes, shown in Table III, can be derived by considering all possible parental mating types and their offspring genotype distributions [] (i. Neale, MC and Cardon, LR: Methodology for Genetic Studies of Twins and Families; in NATO ASI Series D, Behavioural and Social Sciences, vol 67. Dordrecht, Kluwer Academic, 1992).

The effects $\mu(G)$ of genotype G are to displace the phenotypic mean by a, d, and $-a$ for genotypes A_1A_1, A_1A_2, and A_2A_2 respectively, with the raw mean $(2p-1)a + 2p(1-p)d$ then subtracted to preserve the overall phenotypic mean of 0. The relationship between d and a determines the inheritance mode of allele A_1: $d = -a$ for a recessive allele, $+a$ for a dominant allele, and $d = 0$ for an additive allele.

The phenotypic variance contributed by the genotype G can be partitioned into an additive component σ_A^2 and a dominance component σ_D^2, with

$\sigma_A^2 + \sigma_D^2 = 2p(1-p)[a-d(2p-1)]^2 + 4p^2(1-p)^2d^2$.

As will be seen below, this partitioning is important because association tests are sensitive primarily to σ_A^2, not to σ_D^2. Note that σ_A^2 may be much larger than σ_D^2 even when the inheritance is purely dominant or recessive. Remaining genetic and environmental factors contribute a residual variance $\sigma_R^2 = 1 - (\sigma_A^2+\sigma_D^2)$ to the total phenotypic variance.

The probability density of phenotypic values for sib pairs is denoted $f(X_1,X_2)$. It can be expressed as a mixture of 9 conditional densities, one for each possible sib-pair genotype,

$f(X_1,X_2) = \sum_{G_1G_2} f(X_1,X_2|G_1,G_2)P(G_1,G_2)$.

The mean of X_i is $\mu(G_i)$ for sib $i = 1$ or 2; both X_1 and X_2 have residual variance σ_R^2 and residual covariance (due to shared residual genetic and environmental factors) t_R. The total phenotypic correlation t for sib pairs is

$t = t_R + \sigma_A^2/2 + \sigma_D^2/4$

when effects from genotype G are included.

Although X_1 and X_2 are natural coordinates for expressing sib phenotypic values, the correlation between sibs complicates the joint distribution of X_1 and X_2. A simpler joint distribution is obtained by noting that the sum and difference of X_1 and X_2 are completely uncorrelated. These orthogonal coordinates representing sib mean and sib difference are denoted X_+ and X_-, with

$X_\pm = (X_1 \pm X_2)/2$.

The probability distribution in these orthogonal coordinates, $f(X_+,X_-|G_1G_2)$, factors into the product of $f(X_+|G_1,G_2)$ and $f(X_-|G_1,G_2)$, with

$f(X_\pm|G_1,G_2) = (2\pi\sigma_\pm^2)^{-1/2}\exp\{-[X_\pm - \mu_\pm(G_1,G_2)]^2 / 2\sigma_\pm^2\}$, using

$\mu_\pm(G_1,G_2) = [\mu(G_1) \pm \mu(G_2)]/2$ and

$\sigma_\pm^2 = \sigma_R^2(1 \pm t_R)/2$.

It is also convenient to define pair-mean and pair-difference allele frequencies $p_\pm(G_1,G_2)$ as

$p_\pm(G_1,G_2) = (p_{G_1} \pm p_{G_2})/2$.

The variance of the pair-mean and pair-difference variables may be expressed more generally for sib-ships of size s, with genotypic correlation r between any two sibs within a sib-ship, as

$\mathrm{Var}(X_\pm) = \sigma_R^2 \, T_\pm$ and

$\mathrm{Var}(p_\pm) = \sigma_p^2 \, R_\pm$

where

$T_\pm = [1 \pm (s-1)t_R]/s$ and

$R_\pm = [1 \pm (s-1)r]/s$.

The family size s is 2 for sib-pairs, and the genotypic correlation r is 0.5 for full sibs.

In addition to X_1,X_2 and X_+,X_- coordinate systems, we also introduce a Mahalanobis coordinate system. In this metric, a sib-pair is described by a radial coordinate b, which expressed how extreme the pair of phenotypic values is, and an angle φ, which determines whether each sib has a positive or negative phenotypic value. The transformations relating the Mahalanobis variables to the pair-mean and pair-difference variables are

$X_+ = \sigma_+ \, b \sin\varphi$ and

$X_- = \sigma_+ \, b \cos\varphi$.

The probability distribution in Mahalanobis coordinates is

$f(b,\varphi|G_1,G_2) = (2\pi)^{-1} \exp[-(b\sin\varphi - \nu_+)^2/2] \exp[-(b\cos\varphi - \nu_-)^2/2]$ with

$\nu_\pm = \mu_\pm/\sigma_\pm$.

This distribution satisfies

$$\int_0^{2\pi} d\varphi \int_0^{\infty} db \, b \, f(b,\varphi|G_1,G_2) = 1.$$

In the absence of a contribution from the QTL, $f(b,\varphi|G_1,G_2)$ reduces to $(2\pi)^{-1} \exp(-b^2/2)$ and the Mahalanobis probability density is independent of the phase φ. Contour lines of equal probability density in the X_1–X_2 plane are ellipses tilted at 45° with a ratio of major axis to minor axis of $[(1+t)/(1-t)]^{1/2}$.

4.2 Test statistic and pool design

The tests of association described here depend on detecting differences in allele frequency in DNA pooled from individuals chosen from a large repository DNA repository. The allele

frequency in the upper pool, with individuals selected to have higher phenotypic values, is denoted p_U; the allele frequency in the lower pool, selected for lower phenotypic values, is p_L; and the test statistic is $p_U - p_L$, denoted Δp.

The overall repository size is denoted N, composed entirely of either N unrelated individuals or $N/2$ sib pairs. The upper and lower pools each hold n samples, and the pooling fraction ρ is defined as n/N.

For an unrelated population, only one design is described: selecting the n individuals whose phenotypic values are at the upper and lower tails of the distribution, thus defining upper and lower thresholds X_U and X_L. This is termed the unrelated-population design.

A corresponding design for sib pairs is termed unrelated-random. In this design, one sib is chosen, at random, from each sib-ship to generate a population of $N/2$ unrelated individuals. Individuals at the upper and lower tails of this unrelated subset are then selected for pooling. The unrelated-random design for $N/2$ sib pairs with pooling fraction ρ is essentially equivalent to the unrelated-population design for $N/2$ individuals with pooling fraction 2ρ.

A second design selecting only unrelated individuals is termed the Mahalanobis design. The pair-mean X_+ and pair-difference X_- are used to define a Mahalanobis coordinate b according to

$b^2 = 2X_+^2/(1+t) + 2X_-^2/(1-t)$.

The n sib-ships with the largest magnitude b and a positive pair-mean X_+ are identified, and the sibling with the larger phenotypic value is selected for the upper pool. Similarly, the n sib-ships with the largest b and negative pair-mean are identified, and the sibling with the more negative phenotypic value is selected for the lower pool.

Two remaining designs select both members of a sib pair for pooling. The pair-mean design selects each sib-ship as a family unit based on the phenotypic mean of the pair. The $n/2$ pairs at the extreme upper and lower tails of the distribution of phenotypic means for sib-ships, comprising n individuals each, are selected for the upper and lower pools respectively. The upper and lower thresholds are again termed X_U and X_L.

The pair-difference design selects individuals based on the difference of phenotypic values within each sib-ship, or equivalently on the magnitude of within-family phenotypic variance. The n sib-pairs with the greatest within-family variance are identified. Within each pair, the individual with the higher phenotypic value is selected for the upper pool, and the individual with the lower phenotypic value is selected for the lower pool. The threshold for the magnitude of the difference $|X_1-X_2|/2$ for selecting families is termed X_T.

Since the X_+ and X_- are uncorrelated within each family, the results of the pair-mean and pair-difference designs are statistically independent and may be combined to yield a single, potentially more powerful test.

4.3 Test power

Under the null hypothesis H_0, the expectation for p_U and p_L is the population mean allele frequency, and the expectation for the test statistic Δp is zero. Under the alternative hypothesis H_1, the expectation $E_1(\Delta p)$ for Δp is non-zero. The power of a test of Δp depends on the magnitude of $E_1(\Delta p)$ compared to the variation of Δp under H_0 and H_1, and in turn on the variation of p_U and p_L.

Both p_U and p_L follow multinomial distributions defined by the probability that an individual with zero, one, or two copies of allele A_1 is selected for pooling. When the total number of individuals selected for each pool is large and the number of copies of allele A_1 in each pool is also large, the multinomial distribution giving Δp is described accurately by a normal distribution. The variance of Δp under H_0 is denoted σ_0^2/n and the variance under H_1 is denoted σ_1^2/n, where σ_0^2 and σ_1^2 depend on the model parameters and the pooling design. The number of individuals required for type I error rate α and type II error rate β is

$n = (z_\alpha\sigma_0 - z_{1-\beta}\sigma_1)^2/E_1(\Delta p)^2$.

The terms z_α and $z_{1-\beta}$ are normal deviates corresponding to the error rates,

$\Phi(z_\alpha) = 1-\alpha$, and $\Phi(z_{1-\beta}) = \beta$,

where $\Phi(z)$ is the cumulative probability function for the standard normal distribution,

$$\Phi(z) = \int_{-\infty}^{z} dz\,(2\pi)^{-1/2}\exp(-z^2/2).$$

The significance level α is for a one-sided test, which is appropriate for association tests for disease-susceptibility markers. If markers for protective polymorphisms are also sought, the significance for a two-sided test is more appropriate.

The method used here to optimize test designs is to specify the error rates α and β, then calculate the selection criteria that minimize the total repository size N required to achieve these error rates for specific genetic models. The method is outlined below, along with a summary of analytical approximations for the repository sizes required for different population structures and pooling designs. Comparisons of the analytical approximations with essentially exact numerical calculations are found in the Results section, and mathematical details are provided in the Appendix.

To optimize N, a trial value of the fraction ρ is chosen. Next, the threshold phenotypic values that select $n = \rho N$ individuals for each pool are derived from the distribution of phenotypic values. Depending on the pooling design, these threshold values may refer to phenotypes for unrelated individuals, the Mahalanobis measure b, the pair-mean measure X_+, or the pair-difference measure X_-. The threshold values are used to calculate the probabilities $\theta_U(G)$ and $\theta_L(G)$ that an individual selected for the upper and lower pools has a particular genotype G. These probabilities provide the expectation $E_1(\Delta p)$ of Δp under H_1, as well as the variances σ_0^2/n and σ_1^2/n of Δp under H_0 and H_1. Values of Δp larger than $z_\alpha\sigma_0/n^{1/2}$ are significant at level α, and the corresponding power of the test is

$1-\beta = \Phi\{[\rho^{1/2}N^{1/2}E_1(\Delta p)-z_\alpha\sigma_0]/\sigma_1\}$.

Since the terms $E_1(\Delta p)$, σ_0^2, and σ_1^2 depend on ρ but not on N or n, this equation may be inverted to find N as a function of $1-\beta$,

$N = (z_\alpha\sigma_0 - z_{1-\beta}\sigma_1)^2 / \rho E_1(\Delta p)^2$.

Optimization proceeds by a search for the value of ρ giving smallest N.

For complex traits, the total variance $\sigma_A^2+\sigma_D^2$ from any QTL is small, and σ_R^2 is close to 1. This suggests that the displacements a, d, and $-a$ are also small relative to σ_R and motivates a

perturbation expansion of $E_1(\Delta p)$ and σ_1^2 in terms of $\mu(G)/\sigma_R$. The expression for Δp is linear in the expansion parameter, while σ_1^2 is identical to σ_0^2 to first order. Collecting the lowest order terms, the result for the required repository size N is proportional to σ_R^2/σ_A^2. The constant of proportionality depends on the pooling fraction ρ, the phenotypic correlation between sibs, the specified values of α and β, and the pooling design, but not on any properties of the genetic model other than σ_A^2.

In deriving the optimal test designs and estimating the test power, we assume implicitly that there is no measurement error in either the allele frequency p or the allele frequency difference Δp. For the allele frequency p, we show in the Results that either using the mean value $(p_U+p_L)/2$ or measuring the allele frequency on a large pool of individuals unselected for phenotypic value should provide an adequate estimate for p. We also discuss the reduction in power due to measurement error in Δp.

Unrelated design

When a repository contains N unrelated individuals, the analytical approximation for the required repository size, derived in the Appendix, is

$N_{unrelated} = (\rho/2y_\rho^2)\,(z_\alpha - z_{1-\beta})^2\,\sigma_R^2/\sigma_A^2$.

This functions is a minimum at $\rho = 0.27$, with $\rho/2y_\rho^2 = 1.24$.

If the population consists of sib pairs rather than unrelated individuals, an unrelated sub-population of $N/2$ individuals may be constructed by selecting one sib at random from each pair. A direct extension of the above result for unrelated populations yields

$N_{random\text{-}sib} = 2[(2\rho)/2y_{2\rho}^2]\,(z_\alpha - z_{1-\beta})^2\,\sigma_R^2/\sigma_A^2$

for the sib-pair population. The repository size required for sib pairs is twice as large as for unrelated individuals, with a pooling fraction half as large.

Mahalanobis design

The analytical approximation for the number of individuals required for the Mahalanobis design, derived in the Appendix, is

$N_{Mahal} = (2\rho)^{-1}[(2b_\rho/\pi) + \Phi(-b_\rho)/\rho(2\pi)^{1/2}]^{-2}\,[\,R_+/T_+^{1/2} + R_-/T_-^{1/2}\,]^{-2}\,(z_\alpha - z_{1-\beta})^2\sigma_R^2/\sigma_A^2$.

The initial geometrical factor depends only on the pooling fraction. It is minimized at $\rho = 0.188$ with a value of 2.90, yielding

$N_{Mahal} = 2.90\,[\,R_+/T_+^{1/2} + R_-/T_-^{1/2}\,]^{-2}\,(z_\alpha - z_{1-\beta})^2\sigma_R^2/\sigma_A^2$

for this pooling design.

Pair-mean design

The analytical approximation for the repository size required by the pair-mean design is

$N_{pair\text{-}mean} = (s\rho/2y_\rho^2)\,(T_+/R_+)\,(z_\alpha - z_{1-\beta})^2\,\sigma_R^2/\sigma_A^2$,

where $s = 2$ for sib pairs. As with the unrelated design, the factor ρ/y_ρ^2 is optimized with a pooling fraction of 0.27, yielding

$N_{pair\text{-}mean} = 2.47\,(T_+/R_+)\,(z_\alpha - z_{1-\beta})^2\,\sigma_R^2/\sigma_A^2$

for the required repository size.

Pair-difference design

An analytical approximation for the repository size required by the pair-difference design is

$N_{pair\text{-}diff} = (s\rho/2y_\rho^2)(T_-/R_-)(z_\alpha - z_{1-\beta})^2\,\sigma_R^2/\sigma_A^2$.

The factor ρ/y_ρ^2 is minimized with a pooling fraction of 0.27, and

$N_{pair\text{-}diff} = 2.47\,(T_-/R_-)(z_\alpha - z_{1-\beta})^2\,\sigma_R^2/\sigma_A^2$

is the required repository size.

Combined pair-mean and pair-difference design

Because the sib-mean variables X_+ and p_+ are uncorrelated with the pair-difference variables X_- and p_-, the pair-mean and pair-difference estimators are independent and may be combined into a single test. The combined test uses the measured value of $\Delta p_\pm$, where the + and – signs refer to the allele frequency differences found for the pair-mean and pair-difference pools, to obtain an estimator for σ_A/σ_R. The pair-mean and pair-difference estimators, $Q_\pm$, each with expectation σ_A/σ_R, are

$Q_\pm = (T_\pm^{1/2}/R_\pm)(\rho/2y_\rho\sigma_p)\Delta p_\pm$, with

$\mathrm{Var}(Q_\pm) = (s\rho/2y_\rho^2 N)T_\pm/R_\pm$

from the expressions provided in the Appendix for $\mathrm{Var}(\Delta p_\pm)$. These expressions differ by the factor $sR_\pm$ from a similar expression provided by Ollivier et al. (1997) which incorrectly neglected the contribution of family structure to $\mathrm{Var}(\Delta p_\pm)$.

The combined minimum-variance estimator Q having expectation σ_A/σ_R, constructed by weighting the pair-mean and pair-difference estimators according to their inverse variances, is

$Q = (\rho/2y_\rho\sigma_p)\,[(R_+/T_+) + (R_-/T_-)]^{-1}\,(T_+^{-1/2}\Delta p_+ + T_-^{-1/2}\Delta p_-)$, with

$\mathrm{Var}(Q) = (s\rho/2y_\rho^2 N)[(R_+/T_+) + (R_-/T_-)]^{-1}$.

An analytical approximation for the repository size required using the combined estimator is

$N_{\mathrm{comb}} = (s\rho/2y_\rho^2)\,[(R_+/T_+) + (R_-/T_-)]^{-1}\,(z_\alpha - z_{1-\beta})^2\sigma_R^2/\sigma_A^2$.

At the optimal pooling fraction of $\rho = 0.27$, the factor $(s\rho/2y_\rho^2)$ is 2.47. Since the variance of the individual estimators are identical under H_0 and H_1, the repository size for the combined estimator is simply the reciprocal of the sum of the reciprocal repository sizes required for the individual estimators.

4.4 Regression tests

Regression tests requiring individual genotyping provide a benchmark for the efficiency of tests on pooled DNA. A regression test assesses the significance of the regression coefficient m in the model

$X_i = m(p_i - p) + \varepsilon_i$

where i labels an observation, X_i is an observed phenotype with mean 0 and variance 1, p_i is the corresponding observed genotype with mean p, and ε_i is the residual contribution not explained by the model. For N unrelated individuals, the phenotypic and genotypic variables in the regression test are the individual X_i and p_i values. For $N/2$ sib-pairs, they are the pair-mean and pair-difference variables $X_\pm$ and $p_\pm$ for each pair.

The expectation of the regression coefficient m is 0 under H_0 and is

$\mathrm{E}(m) = \sigma_A/\sigma_p$,

under H_1. The variance of the estimator, assumed identical under both hypotheses with negligible error when σ_R^2 is close to 1, is

$\mathrm{Var}(m) = (s/N)\ \mathrm{Var}(\varepsilon_i)/\mathrm{Var}(p_i) = (s/N)(T/R)\sigma_R^2/\sigma_p^2$,

where $s = 1$ for unrelated individuals or 2 for sib-pairs, and $T/R = 1$ for unrelated individuals and $T_\pm/R_\pm$ for pair-mean and pair-difference variables.

The expectation and variance of the test statistic are related to the false-positive rate and power through the equation

$[\mathrm{Var}(m)]^{-1} = (z_\alpha - z_{1-\beta})^2/[\mathrm{E}(m)]^2$.

Substitution into this equation yields the repository size requirement for the regression test,

$N_{\mathrm{regr}} = s(T/R)(z_\alpha - z_{1-\beta})^2\sigma_R^2/\sigma_A^2$.

The combined estimator formed from the pair-mean and pair-difference estimators has a repository size requirement of

$N_{\mathrm{regr}} = s[R_+/T_+ + R_-/T_-]^{-1}(z_\alpha - z_{1-\beta})^2\sigma_R^2/\sigma_A^2$.

4.5 Computational methods

Results for required repository sizes were obtained numerically using computations converged to $1\text{x}10^{-6}$ []. ([ii] Press, WH, Teukolsky, SA, Vetterling, WT, and Flannery, BP: Numerical recipes in C, the art of scientific computing, ed 2. Cambridge, UK, Cambridge University Press, 1997.)

Brent's root-finding algorithm was used to determine the threshold values X_U and X_L for the upper and lower pools for a given pooling design and pooling fraction ρ; Brent's optimization algorithm was then used to find the ρ with maximum power. While the reported results are based on a normal approximation for the allele frequency difference Δp, results were also obtained using the underlying multinomial distribution for the unrelated-population design. The difference between the numerical results for the multinomial and normal distributions was typically less than 1%. The repository size required for the pooling combined estimator was obtained numerically as the reciprocal of the sum of the reciprocal exact sizes required for the pair-mean and pair-difference pooling designs. Using a 750 MHz Pentium III running Linux, the root-finding and minimization for each parameter set required less than 0.01 sec for each design.

To assess the error made by assuming a normal distribution for Δp, we also performed tests in which Δp was calculated exactly according to a multinomial distribution. Results for the required repository size based on the normal distribution were then compared to the repository size based on a multinomial distribution. The two results for N differed by no more than 5% when the number of copies of the minor allele summed over both pools is greater than 60. They differ by approximately 10% when the number of alleles is 10, with the normal distribution underestimating the exact repository size. These differences are not visible on the scale of the figures.

Appendix 4A: Mathematical details

4A.1 Unrelated design

The unrelated design considers a population of N unrelated individuals. Upper and lower thresholds X_U and X_L are defined using

$\rho = \Sigma_G \Phi\{-[X_U-\mu(G)]/\sigma_R\}P(G)$ and

$\rho = \Sigma_G \Phi\{[X_L-\mu(G)]/\sigma_R\}P(G)$,

which may be inverted numerically to find X_U and X_L as functions of ρ. The probability that an individual selected for a pool has genotype G is denoted $\theta_U(G)$ for the upper pool and $\theta_L(G)$ for the lower pool,

$\theta_U(G) = \rho^{-1}\Phi\{-[X_U-\mu(G)]/\sigma_R\}P(G)$ and

$\theta_L(G) = \rho^{-1}\Phi\{[X_L-\mu(G)]/\sigma_R\}P(G)$.

The expected allele frequencies under H_1 are

$E_1(p_U) = \Sigma_G \theta_U(G)\, p_G$ and

$E_1(p_L) = \Sigma_G \theta_L(G)\, p_G$, with

$E_1(\Delta p) = E(p_U) - E(p_L)$.

The variance of the test statistic can be obtained from the moments of a multinomial distribution [] ([iii] Beyer WH (ed): CRC Standard Mathematical Tables, ed 27. Boca Raton, CRC Press, 1984.),

$\sigma_0^2 = 2\ \{\Sigma_G P(G)p_G^2\} - 2p^2 = 2\sigma_p^2$ and

$\sigma_1^2 = \Sigma_G [\theta_U(G) + \theta_L(G)]\, p_G^2 - (p_U^2 + p_L^2)$.

Thus, when ρ is specified, the terms in the expression for the repository size N, $(z_\alpha\sigma_0-z_{1-\beta}\sigma_1)^2/\rho E_1(\Delta p)^2$, may all be calculated numerically, and the optimal ρ is obtained by numerical minimization of N as a function of ρ.

An approximate analytical expression for N may be obtained when σ_R^2 is close to 1 by noting that

$\Phi(z-\delta) = \Phi(z) - y\delta$,

where $y = (2\pi)^{-1/2}\exp\{-z^2/2\}$, is correct to lowest order in the small parameter δ. Using $\mu(G)/\sigma_R$ as the small parameter δ, the phenotypic thresholds are

$X_U = -X_L = -\sigma_R\,\Phi^{-1}(\rho)$, and

the expected difference in allele frequency is

$E(\Delta p) = 2y_\rho[\Sigma_G P(G)\mu(G)p_G]/\rho\sigma_R = 2y_\rho\sigma_p\sigma_A/\rho\sigma_R$,

where $y_\rho = (2\pi)^{-1/2}\exp\{-[\Phi^{-1}(\rho)]^2/2\}$. To the same order of approximation in $\mu(G)/\sigma_R$, both σ_0^2 and σ_1^2 may be replaced with $2\sigma_p^2$. The resulting approximation for the required repository size is

$N_{\text{unrelated}} = (\rho/2y_\rho^2)\,(z_\alpha-z_{1-\beta})^2\,\sigma_R^2/\sigma_A^2$.

The minimum occurs at $\rho = 0.27$ and $y_\rho = 0.33$.

4A.2 Mahalanobis design

For the Mahalanobis design, thresholds b_U and b_L for the radial coordinate are established for the upper and lower pool by solving the following normalization equations:

$$\rho = (1/2)\sum_{G_1,G_2} P(G_1,G_2)\int_0^{\pi} d\varphi \int_{b_U}^{\infty} db\, b\, f(b,\varphi|G_1,G_2) \text{ and}$$

$$\rho = (1/2)\sum_{G_1,G_2} P(G_1,G_2)\int_{\pi}^{2\pi} d\varphi \int_{b_L}^{\infty} db\, b\, f(b,\varphi|G_1,G_2).$$

The factor of (1/2) arises because only one individual is selected from each sib pair. If the radial coordinate b is larger than the threshold value, the phase angle φ determines which sib is selected for which pool: the sibling with genotype G_1 is selected for the upper pool if $0 < \varphi < \pi/2$ and for the lower pool if $\pi < \varphi < 3\pi/2$; the sibling with genotype G_2 is selected for

the upper pool if $\pi/2 < \varphi < \pi$ and for the lower pool if $3\pi/2 < \varphi < 2\pi$. The genotype probabilities $\theta_U(G)$ and $\theta_L(G)$ for the upper and lower pools may be written

$$\theta_U(G) = \rho^{-1} \Sigma_{G'} P(G,G') \int_0^{\pi/2} d\varphi \int_{b_U}^{\infty} db\, b\, f(b,\varphi|G,G') \text{ and}$$

$$\theta_L(G) = \rho^{-1} \Sigma_{G'} P(G,G') \int_{\pi}^{3\pi/2} d\varphi \int_{b_L}^{\infty} db\, b\, f(b,\varphi|G,G'),$$

where symmetry between siblings has allowed the change in integration limits for φ to consider only the regions where sibling 1 is selected. Once ρ is specified, the thresholds for b may be obtained numerically, and $E_1(\Delta p)$ may be obtained from θ_U and θ_L. Numerical results for the required repository size may then be obtained as outlined above for the unrelated design.

An analytic approximation for the repository size requirement may be obtained by noting that

$f(b,\varphi|G_1,G_2) = (2\pi)^{-1} [1 + (bv_+)\cos\varphi + (bv_-)\sin\varphi] \exp(-b^2/2)$

to lowest order in the gene effect $\mu(G)$. The normalization condition leads to the equation

$\rho = (1/4)\exp(-b_\rho^2/2)$,

with $b_U = b_L = b_\rho$ defined in terms of the pooling fraction ρ. The genotype frequencies in the upper and lower pools are

$\theta_{U,L}(G) = P(G) \pm \Sigma_G P(G,G') (v_+ + v_-)[(2b_\rho/\pi) + \Phi(-b_\rho)/\rho(2\pi)^{1/2}]$,

where the upper pool has the + sign and the lower pool the – sign. The expected allele frequencies in the upper and lower pools are

$E(p_{U,L}) = p \pm [(2b_\rho/\pi) + \Phi(-b_\rho)/\rho(2\pi)^{1/2}] [R_+/T_+^{1/2} + R_-/T_-^{1/2}] \sigma_p\sigma_A/\sigma_R$,

where the upper pool has the positive deviation from p and the lower pool the negative deviation. These results are derived using the identities

$$\sum_{G_1,G_2} P(G_1,G_2)\, \mu(G_1)\, p_{G_1} = (1/r) \sum_{G_1,G_2} P(G_1,G_2)\, \mu(G_1)\, p_{G_2} = \sigma_A\sigma_p$$

where r is the genotypic correlation (0.5 for full-sibs). Since $\theta_U(G)+\theta_L(G)$ is $2P(G)$, the variance term σ_1^2 is equal to σ_0^2, and both are equal to $2\sigma_p^2$ because all the pooled individuals are unrelated. The approximate expression for the number of individuals required for the Mahalanobis design is

$N_{\text{Mahalanobis}} = (2\rho)^{-1}[(2b_\rho/\pi) + \Phi(-b_\rho)/\rho(2\pi)^{1/2}]^{-2} [R_+/T_+^{1/2} + R_-/T_-^{1/2}]^{-2} (z_\alpha - z_{1-\beta})^2 \sigma_R^2/\sigma_A^2$.

The minimum occurs at $\rho = 0.188$.

4A.3 Pair-mean design

The fraction ρ of the total population selected according to pair-mean pooling is defined in terms of the upper threshold X_U and the lower threshold X_L as

$$\rho = \sum_{G_1,G_2} P(G_1,G_2)\,\Phi\{-[X_U-\mu_+(G_1,G_2)]/\sigma_+\} = \sum_{G_1,G_2} P(G_1,G_2)\,\Phi\{[X_L-\mu_+(G_1,G_2)]/\sigma_+\}.$$

The genotype distribution describing the individuals selected for each pool follows a multinomial distribution based on sib-pair genotypes rather than individual genotypes, such that

$$1 = \sum_{G_1,G_2} \theta_U(G_1,G_2) = \sum_{G_1,G_2} \theta_L(G_1,G_2),$$

with

$\theta_U(G_1,G_2) = \rho^{-1}\Phi\{-[X_U-\mu(G)]/\sigma_R\}P(G_1,G_2)$ and

$\theta_L(G_1,G_2) = \rho^{-1}\Phi\{[X_L-\mu(G)]/\sigma_R\}P(G_1,G_2)$.

The expected allele frequencies under H_1 are

$$E_1(p_U) = \sum_{G_1,G_2} \theta_U(G_1,G_2)\,p_+(G_1,G_2) \text{ and}$$

$$E_1(p_L) = \sum_{G_1,G_2} \theta_L(G_1,G_2)\,p_+(G_1,G_2), \text{ with}$$

$E_1(\Delta p) = E(p_U) - E(p_L)$

and $p_+(G_1,G_2)$ is the pair-mean allele frequency as defined previously. The terms giving the variance of the test statistic under H_0 and H_1 are

$$\sigma_0^2 = 2s\,\{\sum_{G_1,G_2} P(G_1,G_2)[p_+(G_1,G_2)]^2\} - 2sp^2 = 2sR_+\sigma_p^2 = 3\sigma_p^2 \text{ and}$$

$$\sigma_1^2 = s\{\sum_{G_1,G_2} [\theta_U(G_1,G_2) + \theta_L(G_1,G_2)]\,[p_+(G_1,G_2)]^2\} - s(p_U^2 + p_L^2).$$

The factor $s = 2$ accounts for the family structure, as n/s rather than n measurements of p_+ are used to determine the allele frequency of each pool. The variance under the null hypothesis may be derived directly from the sib-pair genotype frequencies, or more simply by noting that the variance of the mean allele frequency for a sib-pair is $R_+\sigma_p^2$, which is (3/4) of the variance σ_p^2 for an individual. Taking the mean of $n/2$ such terms reduces the variance for each pool by

$n/2$. The total variance is obtained by multiplying by 2 for the number of pools, yielding $3\sigma_p^2$. Given ρ, the pooling thresholds are obtained numerically, then used to calculate $E_1(\Delta p)$ and σ_1^2, yielding a numerical result for the repository size N as a function of ρ.

An analytical approximation follows the same derivation used for the unrelated design, except that individual genotypes are replaced by sib-pair genotypes, and individual phenotypes, phenotype offsets, and allele frequencies are replaced by their pair-mean analogs. The upper and lower pooling thresholds are

$$X_U = -X_L = -\sigma_+ \Phi^{-1}(\rho),$$

and the allele frequency difference between pools is

$$E(\Delta p) = 2y_\rho \left[\sum_{G_1,G_2} P(G_1,G_2)\ \mu_+(G_1,G_2)\, p_+(G_1,G_2) \right] / \rho\sigma_+ = (2y_\rho/\rho)\,(R_+/T_+^{1/2})\sigma_p\sigma_A/\sigma_R,$$

where y_ρ is the height of the standard normal density at $\Phi^{-1}(\rho)$ as before. The contributions of the corresponding low-order terms in σ_1^2 cancel, and the variance of Δp is the same under both hypotheses. The repository size required by the pair-mean design is

$$N_{\text{pair-mean}} = (s\rho/2y_\rho^2)\,(T_+/R_+)\,(z_\alpha - z_{1-\beta})^2\,\sigma_R^2/\sigma_A^2.$$

4A.4 Pair-difference design

Under the pair-difference design, a sib pair is selected if the pair-difference X_- is larger in magnitude than a threshold X_T,

$$2\rho = \sum_{G_1,G_2} P(G_1,G_2)\,\Phi\{[\mu_-(G_1,G_2) - X_T]/\sigma_-\} + \sum_{G_1,G_2} P(G_1,G_2)\,\Phi\{-[\mu_-(G_1,G_2) + X_T]/\sigma_-\}.$$

In the first term, sibling 1 has the higher phenotype and is selected for the upper pool, and sibling 2 is selected for the lower pool. In the second term, the roles of the siblings are reversed. Multinomial distributions are defined as $\theta_U(G_1,G_2)$, the genotype probabilities for sib pairs in which sibling 1 enters the upper pool, and $\theta_L(G_1,G_2)$, when sibling 1 enters the lower pool. For selected pairs,

$$1 = \sum_{G_1,G_2} \{\theta_U(G_1,G_2) + \theta_L(G_1,G_2)\}.$$

This normalization implies that

$$\theta_U(G_1,G_2) = (2\rho)^{-1} P(G_1,G_2)\,\Phi\{[\mu_-(G_1,G_2) - X_T]/\sigma_-\} \text{ and}$$

$\theta_L(G_1,G_2) = (2\rho)^{-1}P(G_1,G_2)\,\Phi\{-[\mu_-(G_1,G_2)+X_T]/\sigma_-\}$.

Due to symmetry, $\theta_U(G_1,G_2)$ and $\theta_L(G_2,G_1)$ are identical. The expected allele frequency difference between pools is

$$E_1(\Delta p) = \sum_{G_1,G_2} 2\theta_U(G_1,G_2)p_-(G_1,G_2) - \sum_{G_1,G_2} 2\theta_L(G_1,G_2)p_-(G_1,G_2);$$

by symmetry, each term contributes $E(\Delta p)/2$. To calculate the variance of Δp, it is important to note that the normalization of θ_U and θ_L to 1/2 implies that the probabilities for a multinomial distribution are $2\theta_U$ and $2\theta_L$, with both θ_U and θ_L equal to $P(G_1,G_2)/2$ under the null hypothesis. The terms giving the variance under the null hypothesis and the alternative hypothesis are

$$\sigma_0^2 = 2s\sum_{G_1,G_2} P(G_1,G_2)p_-^2 = 2sR_-\sigma_p^2 = \sigma_p^2 \text{ and}$$

$$\sigma_1^2 = 2\sum_{G_1,G_2} [2\theta_U(G_1,G_2) + 2\theta_L(G_1,G_2)]p_-^2 - E(\Delta p)^2.$$

The value of σ_0^2 under the null hypothesis may be obtained more simply by noting that the allele frequency difference between two siblings has variance σ_p^2, and the measured allele frequency difference is the mean of n such terms.

The repository size required to detect association may be determined exactly by numeric calculation of the threshold value X_T as a function of the pooling fraction ρ. This value is then used to determine $E(\Delta p)$, σ_0^2, and σ_1^2.

An analytic expression accurate when σ_R^2 is close to 1 may be derived using the same technique as for the previous pooling designs. The analytic estimate for the threshold value is

$X_T = -\sigma_-\Phi^{-1}(\rho)$

and the allele frequency difference is

$$E(\Delta p) = 2y_\rho\left[\sum_{G_1,G_2} P(G_1,G_2)\,\mu_-(G_1,G_2)\,p_-(G_1,G_2)\right]/\rho\sigma_- = (2y_\rho/\rho)\,(R_-/T_-^{1/2})\sigma_p\sigma_A/\sigma_R$$

where y_p is the height of the standard normal density at $\Phi^{-1}(\rho)$. The variance term σ_1^2 equals σ_0^2 to this order of approximation, and the repository size required by the pair-difference design is

$N_{\text{pair-diff}} = (s\rho/2y_\rho^2)(T_-/R_-)(z_\alpha - z_{1-\beta})^2\,\sigma_R^2/\sigma_A^2$.

Example 4.1. C mparisons with individual genotyping

When the effect of a QTL is small and the residual variance σ_R^2 is close to 1, the analytic expressions for repository size requirements are exact. In this limit, we begin by comparing the efficiency of pooled DNA designs relative to individual genotyping.

The repository size requirements of pooled DNA methods are shown in Fig. 9 relative to the corresponding regression tests for the same family structure. Methods plotted are the unrelated, pair-mean, pair-difference, and combined designs, as well as the Mahalanobis design. Except for the Mahalanobis design, the ratio of repository size requirements is independent of all model parameters except for the fraction ρ of individuals whose DNA is pooled. Furthermore, the ratio is independent of family structure for these matched comparisons. The optimal pooling fraction is $\rho = 0.27$. The curves are flat near the minimum, indicating that pooling fractions close to the optimum give near-optimal results. Repository sizes must be increased by 1.24× to attain the same power as would have been achieved with N individual genotypes.

The repository size required for the Mahalanobis design is shown relative to that required for the combined regression test. This ratio depends on the residual phenotypic correlation t_R between siblings, and a typical value $t_R = 0.6$ has been selected for illustrative purposes. The minimum at 0.188 is independent of t_R, and the repository must be 1.55× larger than that for a genotyping study.

In Fig. 10, the performance of the Mahalanobis design relative to the combined regression test for individual genotypes is shown as a function of the residual sibling phenotypic correlation t_R, with the optimal fraction 0.188 used to construct the upper and lower pools. The ratio of repository sizes is roughly 1.5 until the phenotypic correlation rises above 0.6, at which point the repository size requirements for the Mahalanobis design begin to rise more steeply.

Example 4.2 Comparisons between unrelated and sib-pair populations

In Fig. 11, the repository size requirements for association tests using DNA pooled from sib pairs are shown as a function of the residual sibling phenotypic correlation t_R, relative to the repository size required for a test of DNA pooled from unrelated individuals. Ratios larger than 1 indicate that the population of N unrelated individuals is more powerful than a population of $N/2$ sib pairs, while ratios smaller than 1 indicate that the sib-pair population is more powerful. These ratios are derived from the analytical approximations derived for complex traits.

For designs using only 2 pools, a population of unrelated individuals is more powerful than a population of sib pairs except for large values of the sibling phenotypic correlation, $t_R > 0.75$, at which point the Mahalanobis and pair-difference designs become more powerful. Below this phenotypic correlation, the Mahalanobis design is substantially more powerful than the other sib-pair tests; above this correlation, the pair-difference design is only slightly more powerful than the Mahalanobis design.

The slope of the pair-difference repository size requirement is 3× larger than the slope of the pair-mean population requirement. Thus, relative to the pair-mean design, the pair-difference design decreases in power rapidly as t_R falls below 0.5 and increases in power rapidly as t_R rises above 0.5.

The combined 4 pool test using pair-mean and pair-difference pools is uniformly the most powerful sib-pair design for all values of t_R. Its worst-case performance relative to an unrelated population occurs when t_R is $(3^{1/2}+1)/(3^{1/2}-1)$, or 0.2679, where it requires a population 7% larger. The unrelated and sib-pair tests require the same repository size when the phenotypic correlation is 0.5, and the sib-pair test becomes much more powerful for equal repository sizes for larger values of t_R.

Example 4.3 Sensitivity to QTL effect size, allele frequency, and inheritance mode

According to the analytic theory, the necessary size of the study population for pooling tests is inversely proportional to the additive variance contributed by the QTL relative to the residual

phenotypic variance, σ_A^2/σ_R^2, and independent of any remaining parameters of the genetic model. Here we provide exact numerical results to assess the region of validity for the analytical approximations. For these numerical results, the type I error rate α is 5×10^{-8} and the type II error rate β is 0.2 to provide adequate power and an acceptable number of false-positives for a whole-genome scan. For consistency in Figs. 4-6, the unrelated-population design is a dotted line, Mahalanobis is a thin line, pair-mean is dashed, pair-difference is dot-dashed, and the combined estimator sib-combined is a thick line.

A single representative value for the sibling phenotypic correlation t_R was selected for these tests. This correlation is equal to half the genetic heritability plus the shared environmental contribution to the total variance of a complex trait. For cancer, heritability has been estimated at 20% and environmental factors at 7% (Verkasalo et al., 1999); for systolic and diastolic blood pressure, heritabilities are estimated at 10% to 50% and environmental factors at 20% to 40% [,] (Iselius et a.., 1983; Perusse et al., 1989); heritability for cholesterol level is estimated at 70% to 90% (Austin et al., 1987) and environmental factors for serum lipids are estimated 15% [] ([viii] Heller DA, de Faire U, Pedersen NL, Dahlen G, McClearn GE: Genetic and environmental influences on serum lipid levels in twins. N Engl J Med 1993; 328: 1150-6). Additional heritability estimates are 20% to 40% for Type 2 diabetes mellitus (NIDDM) [[ix] Watanabe RM, Valle T, Hauser ER, Ghosh S, Eriksson J, Kohtamaki K, Ehnholm C Ehnholm C, Tuomilehto J, Collins FS, Bergman RN, Boehnke M: Familiality of quantitative metabolic traits in Finnish families with non-insulin-dependent diabetes mellitus. Finland-United States Investigation of NIDDM Genetics (FUSION) Study investigators. Hum Hered 1999; 49: 159-168] and 50% for pulmonary function [[x] Wilk JB, Djousse L, Arnett DK, Rich SS, Province MA, Hunt SC, Crapo RO, Higgins M, Myers RH: Evidence for major genes influencing pulmonary function in the NHLBI family heart study. Genet Epidemiol 2000; 19: 81-94]. These values suggest a range of 0.25 to 0.75 for t_R; we selected $t_R = 0.6$. Choosing a

different value of t_R changes the relative power of different pooling designs, as shown in Fig. 11, but does not alter any conclusions regarding the validity of the analytic theory.

In Fig. 12, the ratio σ_A^2/σ_R^2 is varied over 3 orders of magnitude. The QTL has purely additive inheritance and the minor allele frequency is 0.1. The pooling fraction has been optimized numerically, and linearity in the log-log plot demonstrates validity of the analytic results. Inspection of the results shows that agreement extends almost to $\sigma_A^2/\sigma_R^2 = 1$, where the QTL is responsible for half the phenotypic variance, for all the designs except Mahalanobis. The Mahalanobis design is less powerful than predicted by analytic theory for $\sigma_A^2/\sigma_R^2 > 0.05$. This level of additive variance marks the onset of a major gene effect: carriers of the minor allele separating into a clearly resolved affected population, and the association may be identified by traditional family-based linkage analysis.

The allele frequency difference at the significance threshold, $z_\alpha \sigma_0/n^{1/2}$, is shown in Fig. 12B for the same set of parameters. For the combined design, there are actually two frequency differences, one for the pair-mean pools and another for the pair-difference pools. Only the difference for the pair-difference pools is shown. As the QTL contribution becomes smaller, allele frequency differences must be measured with greater precision. While raw frequency differences of 10% are significant for a major gene ($\sigma_A^2/\sigma_R^2 \sim 0.1$), raw frequencies differences of 3% must be measured with little error to achieve maximum power for a complex trait with $\sigma_A^2/\sigma_R^2 \sim 0.01$.

The sensitivity of the results to both the allele frequency p and the inheritance mode are shown in Figs. 5 and 6. In both of these figures, the pooling fractions are fixed at the limiting values 0.27 for the unrelated-population, pair-mean, pair-difference, and sib-combined designs and at 0.188 for the Mahalanobis design, as would be presumably be done if DNA is pooled once then used repeatedly in a genome-wide screen of markers. In Fig. 13, the allele frequency is varied for a phenotype with dominant inheritance (Fig. 13A), additive inheritance (Fig. 13B), and recessive inheritance (Fig. 13C) of the minor allele. The QTL contribution σ_A^2/σ_R^2 is held fixed at 0.02 for these comparisons. The figures are shown only for the region $p < 0.5$ on a log scale to highlight the behavior for small values of p; additive alleles are symmetric about $p =$

0.5, while dominant major alleles are equivalent to recessive minor alleles and vice versa. It is important to note that the displacements a and d are increased to compensate for a smaller allele frequency p in order to keep σ_A^2 constant and ensure that the limiting behavior for a QTL with small effect is a horizontal line. If the displacements had been held constant, then σ_A^2 would decrease linearly with p and the required repository size would increase as $1/p$.

The repository size is rather insensitive to allele frequency for $p > 0.01$ for dominant and additive inheritance, and for $p > 0.2$ for recessive inheritance, for all but the Mahalanobis design, indicating that the analytic theory is valid in these regions. The repository size required to detect association increases rapidly as the allele frequency decreases below these limits. The Mahalanobis design is more sensitive to the allele frequency than the other designs, losing power rapidly as the allele frequency falls below 0.1 for dominant and additive inheritance and 0.2 for recessive inheritance.

The allele frequency at which the analytic theory loses accuracy may be estimated by noting that the perturbation parameters used to derive the theory are the terms $\mu(G)/\sigma_R$. As the magnitude of these terms approaches 1, or equivalently when the displacements a or d become close to 1, the perturbation theory becomes less reliable. This occurs for $p = \sigma_A^2/8$ under dominant inheritance, $\sigma_A^2/2$ under additive inheritance, and $\sigma_A^{2/3}/2$ for recessive inheritance. In Fig. 13, these values are 0.0025, 0.01, and 0.14, and accurately identify the elbows of the repository size curves.

In Fig. 14, the mode of inheritance is varied while the allele frequency is held fixed at one of three values, $p = 0.5$ (Fig. 14A), 0.25 (Fig. 14B), or 0.1 (Fig. 14C). When $p = 0.5$, the inheritance mode has virtually no effect on the repository size required to detect association. The Mahalanobis design is an exception, with increasing requirements only for strong over-dominance. For $p < 0.5$, the additive variance necessarily vanishes at $d = a/(2p-1)$; when d is close to this value, the population requirements increase dramatically. For $p = 0.25$, this occurs at $d = -2a$. Above this critical value of d, excess A_1A_1 homozygotes are detected in the upper pool; below the critical value, excess A_1A_2 heterozygotes are detected in the lower pool. Although Δp is negative in this region and therefore not significant under a one-sided test of allele A_1, a two-tailed test would yield a significant result. The repository size requirements

are substantially smaller than predicted by analytic theory for this region of strongly over-dominant major alleles.

In the bottom panel, Fig. 14C, the allele frequency is $p = 0.1$ and the critical value of d is $-1.125\,a$. The region of increased population requirements is narrower than in Fig. 14B, and becomes narrower still when p is further reduced, but the general behavior is the same.

Example 4.4 Dependence on type I and type II error rates

We have also investigated the sensitivity of the exact numerical results to specified rates of type I and type II error. In the analytical approximations, this behavior is described entirely by the term $(z_\alpha - z_{1-\beta})^2$, and the optimal pooling fractions are independent of α and β. Comparison with numerical results indicate that the analytical theory is accurate, with no differences seen on the scale of the figures previously presented (results not shown). Using the $(z_\alpha - z_{1-\beta})^2$ scaling and specifying a fixed power of 0.8 ($z_{1-\beta} = -0.84$), for example, a whole-genome scan with $\alpha = 5\times10^{-8}$ ($z_\alpha = 5.33$) requires 1.7× more individuals than a test of 1000 candidate markers with $\alpha = 5\times10^{-5}$ ($z_\alpha = 3.89$) and 6.2× more individuals than a test of a single marker with $\alpha = 0.05$ ($z_\alpha = 1.64$).

Example 4.5 Tests in the presence of population stratification

A marker may show spurious association to a phenotype in the presence of a stratified population. We consider a simple model for stratification in which a population contains at least one sub-population having a mean marker frequency and a mean phenotypic value that both deviate from their respective means in the total population. In individual genotyping, within-family tests such as the transmission disequilibrium test are known to be robust to this type of stratification. Between-family tests, however, may identify spurious associations or miss true associations due to stratification effects.

Tests of pooled DNA in which family members are balanced between pools, such as the pair-difference design, are analogous to within-family tests. The value of σ_A/σ_R estimated from this test is robust to stratification effects. The remaining designs, in particular the pair-mean

design, do not balance family members and are subject to stratification. A suitable test for the presence of stratification, therefore, is to compare the value of σ_A/σ_R estimated separately from the pair-difference and pair-mean pools with the combined estimator in the form of a χ^2 test,

$\chi^2 = \{ [Q_+ - Q]^2 / [s\rho/2y_\rho^2 N][T_+/R_+] \} + \{ [Q_- - Q]^2 / [s\rho/2y_\rho^2 N][T_-/R_-] \}$,

with one degree of freedom. This stratification estimator may also be expressed as

$\chi^2 = [Q_+ - Q_-]^2 / \{ [s\rho/2y_\rho^2 N][T_+/R_+ + T_-/R_-] \}$.

A significant finding for this test, for example at the 0.05 level, indicates that stratification is present and that tests other than the pair-difference test may yield spurious results.

Example 4.6 Allele frequency measurement error

The preceding analysis has assumed that allele frequency measurement errors are negligible. Allele frequencies measured by most technologies, including PCR amplification [[xi] Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A: Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. Gen Res 1998; 8; 111-123], kinetic PCR [[xii] Germer S, Holland MJ, Higuchi R. High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. Gen Res 2000; 10; 258-266], denaturing high performance liquid chromatography [[xiii] Hoogendoorn B, Norton N, Kirov G, Williams N, Hamshere ML, Spurlock G, Austin J, Stephens MK, Buckland PR, Owen MJ, O'Donovan MC: Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. Hum Gen 2000; 107; 488-493], single-strand conformation polymorphism [[xiv] Sasaki T, Tahira T, Suzuki A, Higasa K, Kukita Y, Baba S, Hayashi K: Precise estimation of allele frequencies of single-nucleotide polymorphisms by a quantitative SSCP analysis of pooled DNA. Am J Hum Gen 2001; 68; 214-218], pyrophosphate sequencing [[xv] Alderborn A, Kristofferson A, Hammerling U: Determination of single-nucleotide polymorphisms by real-time pyrophosphate DNA sequencing. Genome Res 2000; 10; 1249-1258], and mass spectrometry [[xvi] Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, Little DP, Strausberg R, Koester H, Cantor CR, Braun A: High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Proc Nat Acad Sci USA 2001; 98; 581-584], are typically reported with

standard errors in the range of 0.01 to 0.02. Assuming a measurement error of 0.01 for p_U and p_L, the resulting error in the population mean estimated as $p = (p_U + p_L)/2$ is 0.007. The measurement error in p affects the calculated repository size N primarily through the terms σ_0^2 and σ_1^2, which are proportional to $p(1-p)$. The relative error in N is proportional to $0.007/p$, less than 10% for minor alleles with frequency greater than 0.07.

The measurement error in Δp, however, has a more deleterious affect on the test power. Again assuming a measurement error of 0.01 for each pool, the measurement error for Δp is $\sqrt{2}$ larger, approximately 0.014. This error can eventually become larger than the sampling error σ_0^2/n for large values of n. In this case, the critical value of Δp depends on the measurement error, not the sampling error. For example, the magnitude of Δp for a two-sided test with significance at the 0.01 level and power 0.95 is $(z_{0.005}-z_{0.95})\times 0.014$, or 0.059 using $z_{0.005} = 2.58$ and $z_{0.95} = -1.64$.

The allele frequency measurement error also sets a lower limit for the effect size that can be detected with a pooled test. For example, using the analytical approximation for Δp for pair-mean pools derived in the Appendix,

$$E_1(\Delta p) = (2y_\rho/\rho)(R_+/T_+^{1/2})\sigma_p\sigma_A/\sigma_R \approx 2.6\times(1+t_R)^{-1/2}p(1-p)|a-(2p-1)d| > 0.059,$$

where the optimized pooling fraction $\rho = 0.27$ is used and the residual variance σ_R^2 is approximated as 1. For a typical phenotypic correlation between sibs, t_R is 0.5, and the effect size that can be detected is

$$|a-(2p-1)d| > 0.028 / p(1-p).$$

For additive inheritance and allele frequency of 0.5, the threshold phenotypic displacement a is 0.11 and the corresponding additive variance is 0.0063. If the minor allele frequency is 0.1, the threshold displacement a is 0.31 and the corresponding additive variance is 0.017.

In the presence of population stratification, the pair-mean pools may give spurious results and pair-difference pools are preferred. Using the expectation for Δp derived in the Appendix for pair-difference pools, we require that

$$E_1(\Delta p) = (2y_\rho/\rho)(R_-/T_-^{1/2})\sigma_p\sigma_A/\sigma_R \approx 0.86\times(1-t_R)^{-1/2}p(1-p)|a-(2p-1)d| > 0.059,$$

where $\rho = 0.27$ and $\sigma_R^2 \approx 1$ as before. For a typical phenotypic correlation between sibs, $t_R = 0.5$, the effect size that can be detected is

$|a-(2p-1)d| > 0.049/p(1-p)$.

For additive inheritance and an allele frequency of 0.5, the critical displacement is 0.20 and the additive variance is 0.02. For a rare minor allele, $p = 0.1$, and additive inheritance, the critical displacement is 0.54, corresponding to an additive variance of 0.05.

5. Model 3

In this model techniques similar to those described in Models 1 and 2 are applied to provide optimized selection criteria for association studies of pooled DNA using the allele frequency difference between pools as a test statistic. It is assumed that samples are drawn from pre-existing population-level DNA repository collected from individuals unselected for any particular phenotype, and that each individual has been measured for a particular phenotype of interest; the goal is to select pools to maximize the power of the test.

Assuming no experimental error in allele frequency measurements on pooled DNA, we determine the selection thresholds that maximize the power to detect association as a function of the frequency, phenotypic displacement, and inheritance mode of a functional polymorphism. The genetic parameters are also described in terms of a genotype relative risk model. Power calculations are then used to derive the repository size required to detect association at specified false-positive and false-negative rates. These calculations are performed at three decreasing levels of accuracy: exact numerical calculations using the true multinomial distribution of the test statistic; numerical calculations based on an approximate normal distribution of the test statistic; and analytical approximations accurate for complex traits where the polymorphism has a small effect on the phenotype.

Results are depicted in terms of the repository sizes required for three types of experimental designs for detecting association with a quantitative phenotype: first, a pooled DNA test using a conventional affected/unaffected classification; second, a pooled DNA test of extreme individuals using optimized selection thresholds; third, individual genotyping of the entire population. We conclude with a discussion of the reduction in power of pooled DNA tests due to experimental measurement error and with suggestions for effective use of pooled DNA tests in practice.

5.1 Computational Methods

The calculation of optimized selection thresholds begins with a model for the genotype-dependent distribution of phenotypic values. A quantitative phenotype, denoted X, is standardized to have unit variance and zero mean. The phenotype is hypothesized to be affected by alleles A_1 and A_2, with frequencies p and $1-p$ respectively, at a particular QTL. The population frequencies $P(G)$ for genotypes $G = A_1A_1$, A_1A_2, and A_2A_2 are assumed to obey Hardy-Weinberg equilibrium. Using standard notation for a variance components model, the effect μ_G of genotype G on phenotype X is a, d and $-a$, for genotypes A_1A_1, A_1A_2, and A_2A_2 respectively. These displacements are each offset by subtracting $(2p-1)a + 2p(1-p)d$ to preserve the overall phenotype mean of zero.

The inheritance mode of the QTL is represented by the displacement d of the heterozygote, for example purely recessive ($d = -a$), additive ($d = 0$), or dominant ($d = +a$) inheritance. The inheritance mode partitions the phenotypic variance due to the QTL into the additive variance σ_A^2 and the dominance variance σ_D^2, with

$$\sigma_A^2 + \sigma_D^2 = 2p(1-p)[a-d(2p-1)]^2 + 4p^2(1-p)^2d^2.$$

This partitioning is important because, as will be seen below, pooled tests are sensitive primarily to the additive component of variance. Note that the additive component may be large even when the inheritance is purely dominant or recessive. The contributions to the phenotype from remaining genetic and environmental factors are assumed to follow a normal distribution with residual variance σ_R^2,

$$\sigma_R^2 = 1-(\sigma_A^2+\sigma_D^2).$$

The genotype-dependent phenotype distributions for each genotype are

$$P(X|G) = (2\pi\sigma_R^2)^{-1/2}\exp[-(X-\mu_G)^2/2\sigma_R^2],$$

normal distributions centered at μ_G with width σ_R. The overall phenotype distribution is the weighted sum of the distributions from each genotype,

$$P(X) = \Sigma_G P(X|G)P(G).$$

For a complex trait in which the QTL makes a small contribution, the three underlying distributions may be unresolved in the observed $P(X)$.

This variance components model may be connected to an equivalent affected/unaffected genotype relative risk model by specifying a threshold phenotypic value X_T that classifies individuals as affected ($X > X_T$) or unaffected ($X < X_T$). The proportion r of the total population that is affected is the overall risk or disease prevalence; the probability that an individual with genotype G is affected, divided by the corresponding probability for an individual with genotype A_2A_2, is the genotype relative risk.

In the tests of pooled DNA considered here, a sample repository of total size N serves as the source of DNA to be selected for one of two pools; not every individual need be selected. The test statistic is the difference in the frequency that a particular allele, here always assumed to be A_1, occurs in the two pools. For a quantitative phenotype, it is natural to specify an upper threshold X_U and a lower threshold X_L as the selection criteria. Individuals with phenotypic values above X_U are selected for the upper pool; individuals with phenotypic values below X_L are selected for the lower pool; and individuals with phenotypic values between X_L and X_U are not pooled at all. The number of individuals selected for each pool is ρN. The fraction ρ expressed in terms of X_U is

$$\rho = \Sigma_G \ \Phi[-(X_U-\mu_G)/\sigma_R]P(G),$$

which is solved numerically to determine X_U. The genotypes of individuals selected by $X > X_U$ follow a multinomial distribution; the probability $\theta_U(G)$ that an individual selected for this pool has genotype G is $\Phi[-(X_U-\mu_G)/\sigma_R]P(G)/\rho$. A multinomial distribution is defined similarly for the lower pool,

$$1 = \Sigma_G \theta_L(G) = \rho^{-1}\Sigma_G \Phi[(X_L-\mu_G)/\sigma_R]P(G),$$

using the lower threshold X_L,

A pooling design based on an affected/unaffected classification is similar: affected individuals are selected for the upper pool; an equivalent number of suitably matched unaffected individuals are selected for the lower pool. The selection thresholds X_U and X_L are identical to the classification threshold X_T. The relative risk for genotype G, expressed in terms of the pooling threshold, is $[\theta_U(G)/P(G)]/[\theta_U(A_2A_2)/P(A_2A_2)]$.

The repository size N required to detect association between genotype G and either the quantitative phenotype X or the affected/unaffected classification depends on the desired type I

error rate α and type II error rate β, the chosen test statistic, and the experimental design, as well as on the underlying genetic model. For a one-sided test of a single marker, $\alpha = 1 - \Phi(z_\alpha)$ and $1-\beta = \Phi(-z_{1-\beta})$, where $\Phi(z)$ is the cumulative probability distribution for standard normal deviate z. For a genome scan, the values $\alpha = 5\times10^{-8}$ ($z_\alpha = 5.33$) and $1-\beta = 0.8$ ($z_{1-\beta} = -0.84$) have been suggested.[5] The null hypothesis is denoted H_0 with all μ_G equal to zero, and the alternative hypothesis is denoted H_1 with at least one non-zero μ_G.

An exact calculation of the repository size required to attain desired error rates for a specified genetic model proceeds as follows. First, a value of the pooling fraction ρ or the disease prevalence r is selected. A trial repository size N is specified, with the number of individuals n selected per pool set to the integer part of ρN or rN. Next, the probability $P_0(i,j,k)$ of selecting i individuals with genotype A_1A_1, j individuals with genotype A_1A_2, and k individuals with genotype A_2A_2, with $i+j+k$ equal to n, is tabulated using the multinomial distribution $P_0(i,j,k) = [n!/(i!j!k!)](p^2)^i(2p-2p^2)^j(1-2p-p^2)^k$.

The frequency of allele A_1 for this pool composition is $(2i+j)/2n$. The probability that two pools selected in this manner differ in frequency by at least Δp is calculated as the sum of $P_0(i,j,k)P_0(i',j',k')$ for all combinations of i,j,k and i',j',k' where $[2(i-i') + (j-j')]/2n \geq \Delta p$.

Significance at level α is attained by increasing Δp until this sum is less than or equal to α. If not even the maximum value $\Delta p = 1$ is sufficient for significance at level α, then a larger value of N is selected for the current value of ρ and the calculation begins anew. Otherwise, multinomial probabilities for pool compositions are calculated under H_1 using $P_U(i,j,k) = [n!/(i!j!k!)]\theta_U(A_1A_1)^i\theta_U(A_1A_2)^j\theta_U(A_2A_2)^k$ for the upper pool and a similar term $P_L(i',j',k')$, with θ_L replacing θ_U, for the lower pool. The probability that the allele frequency difference between the upper and lower pools is at least Δp is obtained as the sum of $P_U(i,j,k)P_L(i',j',k')$ for all compositions i,j,k and i',j',k' where $[2(i-i') + (j-j')]/2n \geq \Delta p$. If this probability is greater than or equal to β, the current N is feasible for type I error α and type II error β and a smaller value for N is attempted. This process continues until the smallest feasible N is found.

For the affected/unaffected design, this procedure is followed for each value of r. For the tail pool design, the smallest feasible value for N is calculated as a function of ρ, and the entire design is optimized by searching for the pooling fraction ρ with the smallest feasible N.

When each pool contains a large number of individuals and many copies of each allele, the distribution of allele frequencies for the pool approaches a normal distribution. The difference in allele frequencies between pools, which continues to serve as the test statistic, approaches a normal distribution as well. The pool sizes required to achieve specified error rates are obtained accurately in this case by approximating the multinomial distributions of allele frequencies as normal distributions. Under H_0, the mean of the test statistic is zero and the variance is $\sigma_0^2/n = p(1-p)/n$, derived by noting that the variance of the frequency difference is twice the variance of the mean for a single pool of n individuals. The allele frequency variance for an individual is $p(1-p)/2$, and averaging over the n individuals reduces the variance by the factor n.

Under H_1, the expected allele frequency difference Δp is

$\Delta p = p_U - p_L = \Sigma_G[\theta_U(G)-\theta_L(G)]\, p_G$,

where the genotype-dependent allele frequency p_G is 1 for $G = A_1A_1$, 0.5 for A_1A_2, and 0 for A_2A_2. The variance is σ_1^2/n, where σ_1^2 is obtained from the multinomial distribution,

$\sigma_1^2 = \Sigma_G\,[\theta_U(G) + \theta_L(G)]\, p_G^2 \; - (p_U^2 + p_L^2)$.

The repository size N required for type I error α and power 1–β is

$n = [z_\alpha \sigma_0 - z_{1-\beta}\sigma_1]^2/\Delta p^2$.

For tail pools, ρ is then varied to find the smallest N.

The normal approximation underestimates the repository size requirement relative to the exact results from the multinomial distribution. When the sum of the alleles in both pools is at least 60, the difference in repository sizes is no greater than 5%. We chose 60 alleles in both pools as the criterion for switching from the multinomial to the normal calculation. Standard algorithms were employed to perform the root search for X_U and X_L, the optimization, and the integration over the tail of a normal distribution.

In the regime of typical complex traits, the effect of any single QTL is small, the residual variance σ_R^2 is nearly 1, and analytical results may be obtained by expanding Δp to second order in the effect size μ_G. This corresponds loosely to a perturbation theory for probability distributions. The Δp expansion in turn requires a Taylor series expansion for $\Phi(z)$,

$\Phi(z-\delta) = \Phi(z) - \delta\,(d/dz)\,\Phi(z) + (1/2)\delta^2\,(d/dz)^2\Phi(z)$,

truncated at second order. The first derivative is

$$(d/dz)\,(2\pi)^{-1/2}\int_{-\infty}^{z} dz'\,\exp(-z^2/2) = (2\pi)^{-1/2}\exp(-z^2/2) \equiv y,$$

where y is the height of the normal distribution at normal deviate z, and the second derivative is

$(d/dz)\,(2\pi)^{-1/2}\exp(-z^2/2) = -yz$.

Summing these terms,

$\Phi(z-\delta) = \Phi(z) - y\delta - (1/2)yz\delta^2$.

Substituting this approximation into the expressions for $\theta(G)$ using $\delta = \mu_G/\sigma_R$ and $z = \Phi^{-1}(1-\rho)$ yields for the tail design

$p_U = p + (y/\rho\sigma_R)\{\Sigma_G P(G) p_G \mu_G\} + (y|z|/2\rho\sigma_R^2)\{\Sigma_G P(G) p_G \mu_G^2\}$ and

$p_L = p - (y/\rho\sigma_R)\{\Sigma_G P(G) p_G \mu_G\} + (y|z|/2\rho\sigma_R^2)\{\Sigma_G P(G) p_G \mu_G^2\}$.

The corresponding expressions for the affected/unaffected pools, with $z = \Phi^{-1}(1-r)$, are

$p_U = p + [y/r\sigma_R]\{\Sigma_G P(G) p_G \mu_G\} + [y|z|/2r\sigma_R^2]\{\Sigma_G P(G) p_G \mu_G^2\}$ and

$p_L = p - [y/(1-r)\sigma_R]\{\Sigma_G P(G) p_G \mu_G\} - [y|z|/2(1-r)\sigma_R^2]\{\Sigma_G P(G) p_G \mu_G^2\}$.

The required sums are

$\Sigma_G P(G) p_G \mu_G = \sigma_A[p(1-p)/2]^{1/2}$, and

$\Sigma_G P(G) p_G \mu_G^2 = (1/2)(1-\sigma_R^2) - 4p^2(1-p)^2 ad + (2p-1)\sigma_D^2/2 \approx \sigma_A^2/2$.

The approximate value $\sigma_A^2/2$ for the second sum neglects the dominance variance and is exact for purely additive inheritance. It serves to simplify the final equations for Δp. Little error is made in the resulting Δp for two reasons: first, even with dominant or recessive inheritance, the additive variance is often larger than the dominance variance; second, this factor is part of a correction term that is already small.

The results for Δp are

$\Delta p = 2^{1/2} y\sigma_0\sigma_A/\rho\sigma_R$, tail pools, and

$\Delta p = [1+ \Phi^{-1}(1-r)\sigma_A/2^{3/2}\sigma_0\sigma_R]\, y\sigma_0\sigma_A/2^{1/2}r(1-r)\sigma_R$, affected/unaffected pools.

To the same order of approximation, σ_1^2 may be equated with σ_0^2, and the number of individuals required per pool is

$n = [z_\alpha - z_{1-\beta}]^2\sigma_0^2/\Delta p^2$.

The preceding three equations lead directly to our main results, Eqs. 1 and 2.

The perturbation theory above is valid when the expansion parameters μ_G/σ_R are small, typically satisfied when $\sigma_A^2/2p(1-p)$ is smaller than 1. In this regime, approximate genotype relative risks may be obtained from the Taylor series expansion for $\theta(G)$. To lowest order, the relative risk for the heterozygote is $1 + (d+a)y/r\sigma_R$, and for the A_1A_1 homozygote is $1 + 2ay/r\sigma_R$. For additive inheritance, $d = 0$, and the relative risk is multiplicative with allele dose when $ay/r\sigma_R$ is small.

If individual genotypes are measured for the N individuals in the population, the regression coefficient b_1 in the regression model

$X = b_1(p_G - p) + \varepsilon$

is a suitable test statistic. The residual contribution ε to the phenotype has mean zero and is uncorrelated with p_G. Under H_0, b_1 has mean zero and variance

$\mathrm{Var}(b_1|H_0) = N^{-1}\,\mathrm{Var}(X)/\mathrm{Var}(p_G) = 1/N[p(1-p)/2]$.

Under H_1, the expected value and the variance of b_1 are

$\mathrm{E}(b_1|H_1) = \mathrm{Cov}(X, p_G)/\mathrm{Var}(X) = \sigma_A[p(1-p)/2]^{1/2}$ and

$\mathrm{Var}(b_1|H_1) = N^{-1}\,\mathrm{Var}(\varepsilon)/\mathrm{Var}(p_G) = \sigma_R^2 / N\,[p(1-p)/2]$.

The repository size required for a one-sided test of b_1 with Type I error α and power $1-\beta$ is

$N = [z_\alpha \mathrm{Var}(b_1|H_0)^{1/2} - z_{1-\beta}\mathrm{Var}(b_1|H_1)^{1/2}]^2/[\mathrm{E}(b_1|H_1)]^2$,

which is presented in simplified form as Eq. 3.

Example 5.1

Two experimental designs are considered using DNA pooled from individuals selected from a pre-existing repository of N samples: affected/unaffected pools, with DNA pooled from n

affected and n unaffected individuals; and tail pools, with DNA pooled from the n most extreme individuals at each tail of the phenotype distribution.

For the affected/unaffected design, the expected number of affected individuals is $n = rN$, and an additional n suitably matched controls are selected from the remainder of the population. An analytical approximation for the repository size is

$$N_{\text{aff/unaff}} = [z_\alpha - z_{1-\beta}]^2 [\sigma_R^2/\sigma_A^2] \cdot 2r(1-r)^2/\{y_r^2 [1 + \Phi^{-1}(1-r)\sigma_A/2^{3/2}\sigma_R p^{1/2}(1-p)^{1/2}]^2\}, \quad \text{(Eq. 1)}$$

where y_r is the height of the standard normal distribution at $\Phi^{-1}(r)$ (see Materials and Methods for derivation). Repository size requirements are minimized with a prevalence of 50%, much larger than values realistic for complex disorders.

The tail pools are parameterized by the fraction $\rho = n/N$ of population N selected for each pool. An analytical approximation for the repository size is

$$N_{\text{tail}} = [z_\alpha - z_{1-\beta}]^2 [\sigma_R^2/\sigma_A^2] \cdot \rho/2y_\rho^2, \quad \text{(Eq. 2)}$$

where y_ρ is the height of the standard normal distribution at $\Phi^{-1}(\rho)$ (see Materials and Methods for derivation). The design is optimized by selecting ρ to minimize $\rho/2y_\rho^2$ and hence N_{tail}. The optimal fraction, 27.03%, is independent of all remaining parameters.

The repository size required to achieve the same error rates using individual genotyping is

$$N_{\text{indiv}} = [z_\alpha - z_{1-\beta}\sigma_R]^2 / \sigma_A^2, \quad \text{(Eq. 3)}$$

based on a regression model of phenotypic value on allele dose (see Materials and Methods for derivation).

Results of the analytical approximations are shown in Fig. 15 with individual genotyping serving as a reference. The tail design, with $\rho = 27\%$ of the population selected for each pool, requires a repository only 1.24× larger than required for individual genotyping. It is also robust to variation in ρ near its optimum, as values from 19% to 37% drop the efficiency no more than 5%. In contrast, for 10% disease prevalence, the affected/unaffected design requires a repository 5.3× larger than that required for individual genotyping and is 4× less efficient than the tail design.

The effect of varying the inheritance mode is shown in Figure 16 for tail pools. In this example, the type I error is 5×10^{-8}, the type II error is 0.2, and the displacement a is 0.25 in units of the phenotypic standard deviation. The heterozygote displacement d varies from $-a$, pure recessive inheritance, to $+a$, pure dominant inheritance. Results are shown for three frequencies of allele A_1: $p = 0.5$, 0.1, and 0.01. Solid lines correspond to exact numerical calculations. In the top panel showing the repository size N, filled circles correspond to analytical approximations, Eq. 1, and are virtually indistinguishable from exact calculations. When $p = 0.5$, A_1 and A_2 have equal frequencies, the additive variance is 0.03125, and the dominance variance is 0 regardless of inheritance mode. Since the population requirements depend primarily on the additive variance, N is independent of the inheritance mode. For allele frequencies below 0.5, the additive variance increases from left to right and the population requirements decreases. The maximum population is required when d equals $a/(2p-1)$, which always falls outside the range depicted. The bottom panel depicts the corresponding values of ρ from the numerical calculations. The optimal pooling fractions fall in a narrow range from 24.5% to 27.5%, close to the analytical approximation of 27.03%.

The effect of varying the additive variance directly, or equivalently the genotype relative risk for an allele of known frequency, is shown in Fig. 17. The top panel of Fig. 17 shows that analytical approximations for N from Eqs. 1 and 2 (solid circles) are nearly indistinguishable from the exact numerical results (dashed and solid lines) when the genotype relative risk is below a factor of 2 to 3. Type I and II error rates are 5×10^{-8} and 0.2 respectively, and the allele frequency is 0.1. The bottom panel shows the corresponding allele frequency difference that must be measured for a significant finding with a test of pooled DNA. For example, alleles carrying a 1.5× heterozygote relative risk, corresponding to an additive variance of 0.01, have a raw frequency difference of 0.04 at significance: the upper pool has an allele frequency of 0.12 and the lower pool a frequency of 0.08. The population size required to achieve significance is 4700, with 1270 individuals selected per pool.

This analysis assumes that allele frequency measurement error is negligible. Allele frequencies measured by most technologies, including PCR amplification, kinetic PCR, denaturing high performance liquid chromatography, single-strand conformation polymorphism, pyrophosphate sequencing, and mass spectrometry, are typically reported with

standard errors in the range of 0.01 to 0.02. Assuming a measurement error of 0.01, the measurement error in the frequency difference is larger by a factor of $\sqrt{2}$, yielding a final error of 0.014. Based on the measurement error, the allele frequency difference of 0.04 in the example above corresponds to a *z*-score of 2.86 and a type I error rate of 0.002.

While this error rate is much larger than the error rate of 5×10^{-8} required for a whole-genome scan, a practical solution is to employ pooled allele frequency measurements as a pre-screen; candidate associations identified by the pre-screen may then be confirmed by individual genotyping of the entire population, or possibly just the extreme tails. Setting a type I error rate for the pre-screen of 0.01 (*z*-score of 2.33), corresponding to an allele frequency difference of 0.033, implies a 100× savings over an equivalent study that does not employ a pre-screen.

This experimental limitation sets a threshold for the effect size that may be identified in a pooled DNA pre-screen. The relationship between the expected value of Δp and the parameters of the genetic model for a SNP with purely additive inheritance is

$\Delta p = 2.44\times[z_\alpha/(z_\alpha - z_{1-\beta})]p(1-p)a$,

where the initial factor of 2.44 arises from the optimized pooled tail design, z_α and $z_{1-\beta}$ correspond to the type I and II errors that would be obtained neglecting measurement error, and a is the phenotypic displacement as before. For use in a pre-screen with a p-value of 0.01 from measurement error alone, $z_\alpha = 2.33$ is reasonable. To retain at least 95% of the true associations, β should be no greater than 0.05, with $z_{1-\beta} = -1.64$. These parameters yield Δp equal to $1.43\times p(1-p)a$, or $p(1-p)a = 0.023$ for the 0.033 frequency difference threshold. For a minor allele frequency of 0.1, this corresponds to a displacement a of 0.26 and an additive variance of 0.012; for allele frequencies of 0.5, the displacement is 0.092 and the additive variance is 0.0042. Thus, the pre-screen retains the power to detect markers with additive variance down to 0.5% to 1.5%, depending on the marker frequency.

References

Abecasis, GR, Cardon, LR, Cookson, WOC (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66: 279-292.

Alderborn A, Kristofferson A, Hammerling U: Determination of single-nucleotide polymorphisms by real-time pyrophosphate DNA sequencing. Genome Res 2000; 10; 1249-1258.

Austin MA, King MC, Bawol RD, Hulley SB, Friedman GD (1987) Risk factors for coronary heart disease in adult female twins. Genetic heritability and shared environmental influences. Am J Epidemiol 125: 308-18.

Barcellos LF, Klitz W, Field LL, Tobias R, Bowcock AM, Wilson R, Nelson MP et al. (1997) Association mapping of disease loci, by use of a pooled DNA genomic screen. Am J Hum Genet 61:734-747.

Beyer WH (ed) (1984) CRC Standard Mathematical Tables, 27th Edition. CRC Press, Boca Raton, FL.

Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, Little DP, Strausberg R, Koester H, Cantor CR, Braun A: High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Proc Nat Acad Sci USA 2001; 98; 581-584.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 1999 Jul;22(3):231-238.

Cardon LR (2000) A Sib-Pair Regression Model of Linkage Disequilibrium for Quantitative Traits. Hum Hered. 50:350-358.

Chandler D. Introduction to Modern Statistical Mechanics. New York: Oxford Univ. Press; 1987

Collins A, Lonjou C, Morton NE (2000) Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci USA 96:15173-15177.

Darvasi A, Soller M (1994) Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. Genetics 138: 1365-1373.

Falconer, DS, MacKay, TFC (1996) Introduction to quantitative genetics. Addison-Wesley, Boston.

Fulker DW, Cherny SS, Cardon LR (1995) Multipoint interval mapping of quantitative trait loci, using sib pairs. Am J Hum Genet 56:1224-1233.

Frank, L (2000) Storm brews over gene bank of Estonian population. Science 286: 1262.

Germer S, Holland MJ, Higuchi R. High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. Gen Res 2000; 10; 258-266.

Goddard KA, Hopkins PJ, Hall JM, Witte JS (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet 66:216-34.

Gu C, Todorov A, Rao DC (1996) Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci. Genet Epidemiol 13:513-533.

Heller DA, de Faire U, Pedersen NL, Dahlen G, McClearn GE (1993) Genetic and environmental influences on serum lipid levels in twins. N Engl J Med 328: 1150-6.

Hoogendoorn B, Norton N, Kirov G, Williams N, Hamshere ML, Spurlock G, Austin J, Stephens MK, Buckland PR, Owen MJ, O'Donovan MC: Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. Hum Gen 2000; 107; 488-493.

Iselius L, Morton NE, Rao DC (1983) Family resemblance for blood pressure. Hum Hered 33: 277-286.

Kruglyak, L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genetics 22: 139-144.

Kruglyak L, Lander ES (1995) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439-454.

Liu, B-H (1997) Statistical Genomics. CRC Press, Boca Raton.

Mathews J, Walker RL (1970) Mathematical methods of physics, second edition. Benjamin/Cummings, London.

Neale, MC and Cardon, LR (1992). Methodology for Genetic Studies of Twins and Families, NATO ASI Series D, Behavioural and Social Sciences, Vol. 67. Kluwer Academic, Dordrecht.

Nilsson A, Rose J (1999) Sweden takes steps to protect tissue banks. Science 286: 894.

Ott J (1999) Analysis of human genetic linkage. Johns Hopkins Univ Pr, Baltimore.

Perusse L, Rice T, Bouchard C, Vogler GP, Rao DC (1989) Cardiovascular risk factors in a French-Canadian population: resolution of genetic and familial environmental effects on blood pressure by using extensive information on environmental correlates. Am J Hum Genet 45: 240-251.

Press, WH, Teukolsky, SA, Vetterling, WT, and Flannery, BP (1997) Numerical Recipes in C, The Art of Scientific Computing, Second Edition. Cambridge University Press, Cambridge, UK.

Rabinow, P (1999) French DNA: Trouble in Purgatory. University of Chicago Press, Chicago.

Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405: 847-856.

Risch NJ, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517.

Risch NJ, Teng J (1998) The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res 8:1273-1288.

Risch NJ, Zhang H (1996) Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. Am J Hum Genet 58:836-843.

Sasaki T, Tahira T, Suzuki A, Higasa K, Kukita Y, Baba S, Hayashi K: Precise estimation of allele frequencies of single-nucleotide polymorphisms by a quantitative SSCP analysis of pooled DNA. Am J Hum Gen 2001; 68; 214-218.

Sham, P (1997) Statistics in Human Genetics. Arnold.

Shaw SH, Carrasquillo MM, Kashuk C, Puffenberger EG, Chakravarti A: Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. Gen Res 1998; 8; 111-123.

Snedecor and Cochran Snedecor GW, Cochran WG. Statistical Methods. 8th Ed. Ames, Iowa: Iowa State University Press; 1989

Verkasalo PK, Kaprio J, Koskenvuo M, Pukkala E (1999) Genetic predisposition, environment and cancer incidence: a nationwide twin study in Finland, 1976-1995. Int J Cancer 83: 743-749.

Watanabe RM, Valle T, Hauser ER, Ghosh S, Eriksson J, Kohtamaki K, Ehnholm C et al. (1999) Familiality of quantitative metabolic traits in Finnish families with non-insulin-dependent diabetes mellitus. Finland-United States Investigation of NIDDM Genetics (FUSION) Study investigators. Hum Hered 49: 159-168.

Wilk JB, Djousse L, Arnett DK, Rich SS, Province MA, Hunt SC, Crapo RO et al. (2000) Evidence for major genes influencing pulmonary function in the NHLBI family heart study. Genet Epidemiol 19: 81-94.

Zhang H, Risch N (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268:1584-1589.

Zhang H, Risch N (1996) Mapping quantitative-trait loci in humans by use of extreme concordant sib pairs: selected sampling by parental phenotypes. Am J Hum Genet 59:951-957.

Tables

Table I. Sib-pair genotype probabilities

Sib Genotype		$P(G_1,G_2)$
G_1	G_2	
A_1A_1	A_1A_1	$p_1^4 + p_1^3p_2 + p_1^2p_2^2/4$
A_1A_1	A_1A_2	$p_1^3p_2 + p_1^2p_2^2/2$
A_1A_1	A_2A_2	$p_1^2p_2^2/4$
A_1A_2	A_1A_1	$p_1^3p_2 + p_1^2p_2^2/2$
A_1A_2	A_1A_2	$p_1^3p_2 + 3p_1^2p_2^2 + p_1p_2^3$
A_1A_2	A_2A_2	$p_1^2p_2^2/2 + p_1p_2^3$
A_2A_2	A_1A_1	$p_1^2p_2^2/4$
A_2A_2	A_1A_2	$p_1^2p_2^2/2 + p_1p_2^3$
A_2A_2	A_2A_2	$p_1^2p_2^2/4 + p_1p_2^3 + p_2^4$

Table II. Pooling Designs

Design Family	Indicators			
Design	I_{U1}	I_{U2}	I_{L1}	I_{L2}
Unrelated				
Unrelated-Random	$H(X_1 - X_U)$	–	$H(X_L - X_1)$	–
Unrelated-Extreme	$H(\lvert X_1\rvert - \lvert X_2\rvert) \times H(X_1 - X_U)$	$H(\lvert X_2\rvert - \lvert X_1\rvert) \times H(X_2 - X_U)$	$H(\lvert X_1\rvert - \lvert X_2\rvert) \times H(X_L - X_1)$	$H(\lvert X_2\rvert - \lvert X_1\rvert) \times H(X_L - X_2)$
Sib-Together				
Concordant	$H(X_1 - X_U) \times H(X_2 - X_U)$	$H(X_1 - X_U) \times H(X_2 - X_U)$	$H(X_L - X_1) \times H(X_L - X_2)$	$H(X_L - X_1) \times H(X_L - X_2)$
Pair-mean	$H(X_+ - X_U)$	$H(X_+ - X_U)$	$H(X_L - X_+)$	$H(X_L - X_+)$
Sib-Apart				
Discordant	$H(X_1 - X_U) \times H(X_L - X_2)$	$H(X_L - X_1) \times H(X_2 - X_U)$	$H(X_L - X_1) \times H(X_2 - X_U)$	$H(X_1 - X_U) \times H(X_L - X_2)$
Pair-difference	$H(\lvert X_-\rvert - X_U) \times H(X_1 - X_2)$	$H(\lvert X_-\rvert - X_U) \times H(X_2 - X_1)$	$H(\lvert X_-\rvert - X_U) \times H(X_2 - X_1)$	$H(\lvert X_-\rvert - X_U) \times H(X_1 - X_2)$

Table III. Sib-pair genotype probabilities

Sib-Pair Genotype		$P(G_1,G_2)$
G_1	G_2	
A_1A_1	A_1A_1	$p^4 + p^3(1-p) + p^2(1-p)^2/4$
A_1A_1	A_1A_2	$p^3(1-p) + p^2(1-p)^2/2$
A_1A_1	A_2A_2	$p^2(1-p)^2/4$
A_1A_2	A_1A_1	$p^3(1-p) + p^2(1-p)^2/2$
A_1A_2	A_1A_2	$p^3(1-p) + 3p^2(1-p)^2 + p(1-p)^3$
A_1A_2	A_2A_2	$p^2(1-p)^2/2 + p(1-p)^3$
A_2A_2	A_1A_1	$p^2(1-p)^2/4$
A_2A_2	A_1A_2	$p^2(1-p)^2/2 + p(1-p)^3$
A_2A_2	A_2A_2	$p^2(1-p)^2/4 + p(1-p)^3 + (1-p)^4$

OTHER EMBODIMENTS

While the invention has been described in conjunction with the detailed description thereof, the foregoing description is intended to illustrate and not limit the scope of the invention, which is defined by the scope of the appended claims. Other aspects, advantages, and modifications are within the scope of the following claims.

What is claimed is:

1. A method for detecting an association in a population of individuals between a genetic locus and a quantitative phenotype, wherein two or more alleles occur at the locus, and wherein the phenotype is expressed using a numerical phenotypic value whose range falls within a first numerical limit and a second numerical limit, the method comprising the steps of

a) obtaining the phenotypic value for each individual in the population;

b) selecting a first subpopulation of individuals having phenotypic values that are higher than a predetermined lower limit and pooling DNA from the individuals in the first subpopulation to provide an upper pool;

c) selecting a second subpopulation of individuals having phenotypic values that are lower than a predetermined upper limit and pooling DNA from the individuals in the second subpopulation to provide a lower pool;

d) for one or more genetic loci, measuring the difference in frequency of occurrence of a specified allele between the upper pool and the lower pool; and

e) determining that an association exists if the allele frequency difference between the pools is larger than a predetermined value.

2. The method described in claim 1 wherein the lower limit and the upper limit are chosen such that, for a specified false-positive rate, the frequency of occurrence of false-negative errors is minimized.

3. The method described in claim 1 wherein the population comprises unrelated individuals.

4. The method described in claim 1 wherein the population comprises related individuals.

5. The method described in claim 3 wherein the predetermined lower limit is set so that the upper pool includes the highest 35% of the population and the predetermined upper limit is set so that the lower pool includes the lowest 35% of the population.

6. The method described in claim 3 wherein the predetermined lower limit is set so that the upper pool includes the highest 30% of the population and the predetermined upper limit is set so that the lower pool includes the lowest 30% of the population.

7. The method described in claim 3 wherein the predetermined lower limit is set so that the upper pool includes the highest 27% of the population and the predetermined upper limit is set so that the lower pool includes the lowest 27% of the population.

8. The method described in claim 2 wherein the individuals in the population are sibling pairs and each pair is ranked according to the phenotypic values of the siblings in each pair, and either (i) both members of the sibling pair are selected for the upper pool; (ii) both members of the sibling pair are selected for the lower pool; or (iii) neither member of the sibling pair is selected.

9. The method described in claim 8 wherein each sibling pair is ranked according to a mean value of the phenotypic values of the siblings in each pair, and wherein both members of the sibling pair are in the same pool.

10. The method described in claim 8 wherein the phenotypic values of the siblings in each pair are both above a predetermined lower limit or both below a predetermined upper limit.

11. The method described in claim 8 wherein the predetermined lower limit is set so that the upper pool includes the pairs with the highest 10% of the mean values in the population and the predetermined upper limit is set so that the lower pool includes the lowest 10% of the mean values in the population.

12. The method described in claim 8 wherein the predetermined lower limit is set so that the upper pool includes the pairs with the highest 15% of the mean values in the population and the predetermined upper limit is set so that the lower pool includes the lowest 15% of the mean values in the population.

13. The method described in claim 8 wherein the predetermined lower limit is set so that the upper pool includes the pairs with the highest 20% of the mean values in the population and the predetermined upper limit is set so that the lower pool includes the lowest 20% of the mean values in the population.

14. The method described in claim 8 wherein the predetermined lower limit is set so that the upper pool includes the pairs with the highest 25% of the mean values in the population and the predetermined upper limit is set so that the lower pool includes the lowest 25% of the mean values in the population.

15. The method described in claim 8 wherein the predetermined lower limit is set so that the upper pool includes the pairs with the highest 27% of the mean values in the population and the predetermined upper limit is set so that the lower pool includes the lowest 27% of the mean values in the population.

16. The method described in claim 2 wherein all individuals in the population are members of sibling pairs, and either (i) one member of a sibling pair is selected for the upper pool and the second member of the sibling pair is selected for the lower pool; or (ii) neither member of a sibling pair is selected.

17. The method described in claim 17 wherein the sibling pairs are ranked by the absolute magnitude of the difference in phenotypic value for the siblings within each pair, the percent of pairs with the greatest difference are identified, and the siblings in each pair are distributed such that the sibling with the high phenotypic value is selected for the upper pool and the sibling with the low phenotypic value is selected for the lower pool.

18. The method described in claim 17 wherein the phenotypic value of one member of the sibling pair is above a predetermined lower limit and the phenotypic value of the second member of the sibling pair is below a predetermined upper limit.

19. The method described in claim 17 wherein the percent of pairs is 80% and the distribution provides 10% of the population in each pool.

20. The method described in claim 17 wherein the percent of pairs is 70% and the distribution provides 15% of the population in each pool.

21. The method described in claim 17 wherein the percent of pairs is 60% and the distribution provides 20% of the population in each pool.

22. The method described in claim 17 wherein the percent of pairs is 50% and the distribution provides 25% of the population in each pool.

23. The method described in claim 17 wherein the percent of pairs is 54% and the distribution provides 27% of the population in each pool.

24. The method described in claim 2 wherein the individuals in the population are sibling pairs and the results obtained by performing the methods described in claims 7 and 15 are combined.

25. The method described in claim 3 wherein the population of unrelated individuals are provided by a process comprising the steps of:

a) providing a population of sibling pairs; and

b) selecting only one member of a sibling pair to be included in the population of unrelated individuals.

26. The method described in claim 25 further comprising the steps of :

a) calculating the overall mean of the phenotypic values in the population;

b) subtracting the mean from each phenotypic value;

c) ranking each sibling pair according to the result of the calculation conducted according to

$(\text{pair-mean})^2/(\text{variance of pair-mean}) + (\text{pair-difference})^2/(\text{variance of pair difference})$

to provide the Mahalanobis rank;

d) identifying a more extreme sibling from each sibling pair as the member of the pair having a greater magnitude of the phenotypic value; and

e) from sibling pairs having extreme Mahalanobis ranks constructing pools using the sibling of the pair having the more extreme phenotypic value.

27. The method described in claim 25, further comprising the steps of:

a) calculating the overall mean of the phenotypic values in the population; and

b) selecting that member of each sibling pair having a phenotypic value such that the absolute value of the difference between the individual's phenotypic value and the overall mean is greater than the difference for the other individual in the pair,

thereby providing a population of unrelated individuals.

28. The method described in claim 25 further comprising the steps of:

a) rank ordering the members of the population of sibling pairs to generate a list wherein the rank order of each member of a sibling pair is obtained as the smaller of:

i) the distance from the first member on the list and

ii) the distance from the last member on the list; and

b) selecting that member of each sibling pair having a lower ranking;

thereby providing a population of unrelated individuals.

29. The method described in claim 25 further comprising the steps of:

a) rank ordering the members of the population of sibling pairs to generate a list wherein the rank order of each member of a sibling pair is obtained as the distance from the phenotype mean; and

b) selecting that member of each sibling pair having a lower ranking;

thereby providing a population of unrelated individuals.

30. The method described in claim 1 wherein the population includes individuals who may be classified into classes.

31. The method described in claim 30 wherein the classes are based on an age group, gender, race or ethnic origin.

32. The method described in claim 31 wherein all the members of a class are included in the pools.

33. The method described in claim 1 for determining the genetic basis of disease predisposition.

34. The method described in claim 33, wherein the genetic locus which is analyzed for determining the genetic basis of disease predisposition contains a single nucleotide polymorphism.